Application of Stochastic Processes in Nonparametric Bayes

by

Yingjian Wang

Department of Electrical and Computer Engineering
Duke University

Date: ______________________

Approved:

____
Lawrence Carin, Supervisor

____
Loren Nolte

____
Galen Reeves

____
Guillermo Sapiro

____
Robert Wolpert

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Electrical and Computer Engineering
in the Graduate School of Duke University
2014

Abstract

Application of Stochastic Processes in Nonparametric Bayes

by

Yingjian Wang

Department of Electrical and Computer Engineering
Duke University

Date: ______________________

Approved:

____
Lawrence Carin, Supervisor

____
Loren Nolte

____
Galen Reeves

____
Guillermo Sapiro

____
Robert Wolpert

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering
in the Graduate School of Duke University
2014

Abstract

This thesis presents theoretical studies of some stochastic processes and their applications in the Bayesian nonparametric methods. The stochastic processes discussed in the thesis are mainly the ones with independent increments - the Lévy processes. We develop new representations for the Lévy measures of two representative examples of the Lévy processes, the beta and gamma processes. These representations are manifested in terms of an infinite sum of well-behaved (proper) beta and gamma distributions, with the truncation and posterior analyses provided. The decompositions provide new insights into the beta and gamma processes (and their generalizations), and we demonstrate how the proposed representation unifies some properties of the two, as these are of increasing importance in machine learning.

Next a new Lévy process is proposed for an uncountable collection of covariate-dependent feature-learning measures; the process is called the kernel beta process. Available covariates are handled efficiently via the kernel construction, with covariates assumed observed with each data sample ("customer"), and latent covariates learned for each feature ("dish"). The dependencies among the data are represented with the covariate-parameterized kernel function. The beta process is recovered as a limiting case of the kernel beta process. An efficient Gibbs sampler is developed for computations, and state-of-the-art results are presented for image processing and music analysis tasks.

Last is a non-Lévy process example of the multiplicative gamma process applied in

the low-rank representation of tensors. The multiplicative gamma process is applied along the super-diagonal of tensors in the rank decomposition, with its shrinkage property nonparametrically learns the rank from the multiway data. This model is constructed as conjugate for the continuous multiway data case. For the non-conjugate binary multiway data, the Pólya-Gamma auxiliary variable is sampled to elicit closed-form Gibbs sampling updates. This rank decomposition of tensors driven by the multiplicative gamma process yields state-of-art performance on various synthetic and benchmark real-world datasets, with desirable model scalability.

Contents

Abstract **iv**

List of Tables **x**

List of Figures **xi**

List of Abbreviations and Symbols **xiii**

Acknowledgements **xv**

1 Introduction **1**

2 Lévy Process and Completely Random Measure **6**

2.1 Lévy process 6

2.1.1 Definition of Lévy process 6

2.1.2 Pure-jump nondecreasing Lévy process and its underlying Poisson process 7

2.2 Completely random measure 8

2.2.1 Definition of completely random measure 8

2.2.2 Lévy measure decomposition 8

2.3 Beta process 9

2.4 Gamma process 10

3 Lévy Measure Decompositions for the Beta and Gamma Processes **12**

3.1 Lévy measure decomposition for the beta process 13

3.1.1 Lévy measure decomposition 13

3.1.2 The Lévy process B_k 14

3.1.3 Simulating the beta process 15

3.1.4 Truncation analysis 17

3.1.5 Posterior estimation 19

3.1.6 Feature learning experiment with Lévy measure decomposition of beta process 21

3.1.7 Relating the IBP and beta process 21

3.2 Lévy measure decomposition for gamma process 22

3.2.1 Lévy measure decomposition 22

3.2.2 Lévy processes Γ_k and Γ_{kh} 24

3.2.3 Simulation of gamma process 24

3.2.4 Truncation analysis 25

3.2.5 Posterior estimation 26

3.2.6 Generalized gamma process and symmetric gamma process . . 27

4 Kernel Beta Process **28**

4.1 Kernel Beta Process 30

4.1.1 Review of beta and Bernoulli processes 30

4.1.2 Covariate-dependent Lévy process 31

4.1.3 Characteristic function of the kernel beta process 32

4.1.4 Relationship to the beta-Bernoulli process 33

4.1.5 Properties of $\mathcal{B}$ 34

4.2 Applications 35

4.2.1 Model construction 35

4.2.2 Extensions 36

4.2.3 Inference 37

4.3 Experiments 39
4.3.1 Hyperparameter settings 39
4.3.2 Music analysis 39
4.3.3 Image interpolation and denoising 41
4.4 Summary 43

5 Scalable Bayesian Low-Rank Tensor Representation 45

5.1 Low-Rank Tensor Decomposition 45
5.1.1 CP Decomposition of Tensor 46
5.1.2 Rank Specification 47
5.1.3 CP Decomposition with MGP 48
5.2 Model and Inference 50
5.2.1 Model Description 50
5.2.2 Inference via Gibbs Sampling 51
5.2.3 Computational Complexity 54
5.3 Related Work 55
5.4 Experiments 56
5.4.1 Low-rank Tensor Completion: Continuous Data 58
5.4.2 Low-rank Tensor Completion: Binary Data 59
5.4.3 Binary Classification with Extracted Factors 61
5.4.4 Image Inpainting 62
5.4.5 Scalability 62
5.5 Summary 64

A Lévy Measure Decomposition 65

A.1 Lévy measure decomposing of beta process 65
A.2 Expectation of B_k 65

A.3 Variance of B_k 66

A.4 Truncation analysis of beta process 68

A.5 Limit of the Lévy measure of IBP 68

A.6 Lévy measure decomposing of gamma process 69

A.7 The expectation of Γ_k and Γ_{kh} 70

A.8 The variance of Γ_k and Γ_{kh} 70

B Kernel beta process **72**

B.1 Proof of Theorem 1 72

B.2 Properties of the KBP 74

Bibliography **76**

Biography **81**

List of Tables

4.1 Comparison of BP and KBP for interpolating images with pixels missing uniformly at random, using standard image-processing images. The top and bottom rows of each cell show results of BP and KBP, respectively. Results are shown when 20%, 30% and 50% of the pixels are observed, selected uniformly at random. 43

5.1 Binary classification using factors learned from tensor decomposition . . . 62

5.2 Image Inpainting: Reconstruction errors (MSE) on different amounts (90%, 80%, and 50%) of missing pixels 63

List of Figures

2.1 Beta process: **Top row:** beta process with a Gaussian base measure. **Bottom row:** 100 independent Bernoulli processes with the beta process as the prior. 10

3.1 Simulation of the truncation errors of the beta process decomposition presented in Theorem 1. (a) Comparison of the truncation errors by round yielded by simulating the beta process decomposition with the theoretical analysis. (b) Comparison of the truncation errors by point yielded by simulating the beta process decomposition and stick-breaking beta process, with different c and γ. 18

3.2 Comparison of the feature dictionary elements learned in image in-painting. (a) Features learned via IBP prior. (b) Features learned via the beta process decomposition. 22

4.1 (a) MFCCs features used in music analysis, where the horizontal axis corresponds to time, for "A Day in the Life". Based on the Gibbs collection samples: (b) frequency on number of *unique* dictionary elements, and (c) *total* number of dictionary elements. 40

4.2 Inference of relationships in music as a function of time, as computed via a correlation of the dictionary-usage weights, for (a) and (b), and based upon state usage in an HMM, for (c). Results are shown for "A Day in the Life." The results in (c) are from (Ren et al., 2010), as a courtesy from the authors of that paper. (a) KBP-FA, (b) BP-FA, (c) dHDP-HMM 41

4.3 Denoising Result: the first column shows the noisy images (PSNR is 15.56 dB for Peppers and 17.54 dB for House); the second and third column shows the results inferred from the BP-FA model (PSNR is 16.31 dB for Peppers and 17.95 dB for House), with the dictionary elements shown in column two and the reconstruction in column three; the fourth and fifth columns show results from BP-FA+ (PSNR is 23.06 dB for Peppers and 26.71 dB for House); the sixth and seventh column shows the results of the KBP-FA+ (PSNR is 27.37 dB for Peppers and 34.89 dB for House). In each case the dictionaries are ordered based on their frequency of usage, starting from top-left. 44

5.1 The CP decomposition of tensors (a three-mode tensor shown for illustration). 47

5.2 Continuous Data: MSE 58

5.3 Empirical distribution of the inferred rank by MGP-CP^a run with 90% and 50% missing data (starting with $R = 1$) 59

5.4 Binary Data: AUC Scores 61

5.5 Image Inpainting: **Top row:** Corrupted images with 90%, 80%, and 50% pixels missing. **Bottom row:** Reconstructed. 63

5.6 Linear scalability on a large-scale but sparse tensor 64

List of Abbreviations and Symbols

Abbreviations

AIC	Akaike information criterion.
ARD	Automatic Relevance Determination.
ARD-CP	ARD based CP.
AUC	Area under curve.
BCP	Bayesian CP decomposition.
BIC	Bayesian information criterion.
BP	Beta process.
BPd	Beta process decomposition.
BPFA	Beta process factor analysis.
BPsb	Stick-breaking beta process.
CP	CANDECOMP/PARAFAC decomposition.
CRM	Completely random measure.
dIBP	Dependent Indian buffet process.
dHBP	Dependent hierarchical beta process.
dHDP	Dependent hierarchical Dirichlet process.
EEG	Electroencephalography data.
FA	Factor analysis.
FIA	Flow injection analysis.
GP	Gaussian process.

HDP	Hierarchical Dirichelet process.
HMM	Hidden Markov model.
IBP	Indian buffet process.
InfTucker^{tp}	Infinite Tucker Decomposition based on t process.
KBP	Kernel beta process.
KBP-FA	Kernel beta process factor analysis.
KBP-FA+	Augmented kernel beta process factor analysis.
LGM	Latent Gaussian model.
MAP	Maximum a posteriori estimation.
MCMC	Markov chain Monte Carlo.
MFCC	Mel frequency cepstral coefficients.
MGP	Multiplicative gamma process.
MGP-CP	CP tensor decomposition driven by MGP.
MGP-CP^{a}	Adaptive MGP-CP model.
MGP-CP^{t}	Truncated MGP-CP model.
MP	Multinomial process.
MSE	Mean-squared-error.
PG	Pólya-Gamma distribution.
PSDTF	Positive semidefinite tensor factorization.
PSNR	Peak signal-to-noise ratio.
RESCAL	Relational learning approach with tensor factorization.
SVM	Support vector machine.
WGN	White Gaussian noise.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor, Prof. Lawrence Carin, who has been a great mentor for me. Thanks for your guidance, patience, passion, and immense helps in the journey of my Ph.D. My every step forward at Duke is with your accompany.

I would also like to thank the rest of my Ph.D. committee, Prof. Loren Nolte, Prof. Galen Reeves, Prof. Guillermo Sapiro, and Prof. Robert Wolpert, for your encouragement, questions, and insightful comments. For Prof. Nolte, our class ECE 281 - “Random Signal Processing” and my two terms TA for the class are exceptionally happy time for me. For Prof. Wolpert, the class STA 205 -“Measure Theory” is my most enjoyable class at Duke.

Last to my family - you are always here with me, and I am always with you.

1

Introduction

If the discipline of *Machine Learning* can be regarded as a giant tree, then the *Bayesian Nonparametric Methods* constitute one of the most robust trunks to support its growing. The Bayesian nonparametric methods provide more freedom in model construction in the real-world data analysis, and yield more intuitive and fitting interpretation in the information retrieving from the data (Hjort et al., 2010). The most prominent distinction of nonparametric methods relative to parametric approaches is the utilization of stochastic *processes* rather than probability *distributions*. For example, a Gaussian process (Rasmussen and Williams, 2006) may be employed to nonparametrically represent general smooth functions on a continuous space of covariates (*e.g.*, time). Recently the idea of nonparametric methods has extended to feature learning and data clustering, with interest respectively in the beta-Bernoulli process (Thibaux and Jordan, 2007a) and the Dirichlet process (Ferguson, 1973). In such processes the nonparametric aspect concerns the number of features/clusters, which are allowed to be unbounded ("infinite"), permitting the model to adapt the number of these entities as the given and future data indicate. The increasing importance of these models in machine learning warrants a detailed theoretical analysis of their properties, as well as simple constructions for their implementation. In this thesis, our focus is on the application of stochastic processes in the Bayesian nonparametric methods, as it is the main thread throughout the research of the author's PhD study.

In Chapter 2, we review an important category of stochastic processes, the Lévy processes (Sato, 1999), which has been widely used in nonparametric methods. We also show the relation between Lévy processes and the completely random measures (Kingman, 1967; Jordan, 2009), since a family of Lévy processes, the pure-jump nondecreasing Lévy processes, also fit into the category of the completely random measure proposed by Kingman (Kingman, 1967). As two representative examples of the Lévy process, the beta process and gamma are discussed with their Lévy measures presented. The beta process (Hjort, 1990) is an example of a Lévy process, which is applied in nonparametric feature learning. The gamma process falls in this family as well, with its normalization the well-known Dirichlet process. Hierarchical forms of such models have become increasingly popular in machine learning (Teh et al., 2006; Teh, 2006; Thibaux and Jordan, 2007a), as have nested models (Blei et al., 2010), and models that introduce covariate dependence (MacEachern, 1999; Williamson et al., 2010; Lin et al., 2010).

As a consequence of the important role these models are playing in machine learning, there is a need for the study of the properties of Lévy processes. As examples of such work, (Thibaux and Jordan, 2007a) and (Paisley et al., 2010) present explicit constructions for generating the beta process, (Teh et al., 2007) derives a construction for the Indian buffet process parallel to the stick-breaking construction of the Dirichlet process (Sethuraman, 1994a), and (Thibaux, 2008) obtains a construction for the gamma process under the gamma-Poisson context. Apart from these specialized construction methods, in (Kingman, 1967) a general construction method for completely random measures is proposed, by first decomposing it into a sum of a countable number of σ-finite measures, and then superposing the Poisson processes according to these sub-measures. By regarding the completely random measure as a Lévy process, this method corresponds to decomposing the Lévy measure, which provides clarity of theoretical properties and simplicity in practical implementation. However this Lévy measure decomposition method has not yet come into wide use in machine learning and statistics, probably due to the nonexistence of a universal construction of the measure decomposition. In Chapter 3 we focus on decomposition representation of beta process and gamma process, and provide with theoretical analyses.

The kernel beta process (KBP) discussed in Chapter 4 is proposed under the motivation to represent the possible dependencies among the real-world data in the feature learning task. Feature learning is an important problem in statistics and machine learning, characterized by the goal of (typically) inferring a low-dimensional set of features for representation of high-dimensional data. It is desirable to perform such analysis in a nonparametric manner, such that the number of features may be learned, rather than *a priori* set. A powerful tool for such learning is the Indian buffet process (IBP) (Griffiths and Ghahramani, 2005), in which the data samples serve as "customers", and the potential features serve as "dishes". It has recently been demonstrated that the IBP corresponds to a marginalization of a beta-Bernoulli process (Thibaux and Jordan, 2007b). The IBP and beta-Bernoulli constructions have found significant utility in factor analysis (Knowles and Ghahramani, 2007; Zhou et al., 2009), in which one wishes to infer the number of factors needed to represent data of interest. The beta process was developed originally by Hjort (Hjort, 1990) as a Lévy process prior for "hazard measures", and was recently extended for use in feature learning (Thibaux and Jordan, 2007b), we therefore here refer to it as a "feature-learning measure." And KBP yields an uncountable number of covariate-dependent feature-learning measures, with the beta process a special case. With the

KBP, the dependencies among the real-world data are represented with the covariate-parameterized kernel functions.

In Chapter 5, we discuss the application of a stochastic process other than the Lévy process, the multiplication gamma process, in the low-rank representation of tensors. For the analysis of multiway data, the probabilistic tensor decomposition methods (Chu and Ghahramani, 2009), in particular Bayesian tensor decomposition methods (Xu et al., 2013; Xiong et al., 2010) are naturally appealing since they provide a principled mechanism for dealing with missing data, allow analysis of diverse data types (continuous, binary, ordinal, etc.) using suitable likelihood models, and make it possible to quantify the uncertainty in the parameter estimates and the predictions (when dealing with missing data). Unfortunately, these methods require that the rank of the decomposition is specified prior to the analysis. The rank-estimation problem is further confounded in the case of tensor data for which rank determination is known to be an NP-hard problem (Hastad, 1990). Finally, scalability is another concern when applying these methods. Inference via MCMC or variational methods can be slow as the tensor size becomes large (in the number of observed entries, in the number/dimensions of tensor modes, or in all of these).

Motivated by these, we present a flexible and scalable nonparametric Bayesian tensor decomposition method for analyzing multiway tensor data. Our method has the following key properties: (1) The tensor rank does not have to be specified beforehand and is learned *adaptively* from the data in a principled way using the theoretically motivated multiplicative gamma process prior (Bhattacharya and Dunson, 2011) on the elements of the core diagonal tensor in the CANDECOMP/PARAFAC (CP) low-rank decomposition of tensors (Kolda and Bader, 2009); (2) Both continuous and binary datasets can be analyzed using a fully Bayesian framework, via simple closed-form Gibbs sampling updates; (3) Inference scales linearly with the number of observed entries in the tensor, which makes inference highly scalable for large but sparsely observed multiway datasets, commonly encountered in application domains such as multirelational networks, recommender systems, etc. Even on *non-sparse* tensors, our framework, capable of dealing with large amounts of missing data, allows us to use a very small fraction of the entire data while achieving reconstruction quality that is close to using the complete data (our experimental results on tasks such as image inpainting corroborate this). Our framework is therefore also scalable for analyzing large-scale *dense* tensors.

The thesis is organized as follows: first in Chapter 2 we review the definition and properties of Lévy processes and completely random measures, with the discussion of the relationship between them, and also two representative examples of Lévy processes, the beta process and gamma process. Next in Chapter 3 we present a new representation method for Lévy processes by following the Lévy measure decomposition principle. Then in Chapter 4 we discuss a new type of Lévy processes designed to describe the dependencies among the data in feature learning tasks, the kernel beta process. Last in Chapter 5, we apply the multiplicative gamma process in the CANDECOMP/PARAFAC (CP) decomposition of tensors for the low-rank representation of tensors, and show its applications in multiway data inference and other related tasks.

2

Lévy Process and Completely Random Measure

Lévy processes (Sato, 1999) and completely random measures (Kingman, 1967) are two closely related concepts, as they both require the independence property. Specifically, some Lévy processes can be regarded as completely random measures. In this section brief reviews and connections are presented for these two important concepts.

2.1 Lévy process

2.1.1 Definition of Lévy process

A Lévy process $X(\omega)$ is a stochastic process with independent increments on a measure space $(\Omega, \mathcal{F})$. Ω is usually taken to be one-dimensional, frequently to represent a stochastic process with variation over time. A stochastic process $X(\omega)$ is a Lévy process if it satisfies the three following conditions (Applebaum, 2009):

1. $X(\varnothing) = 0$ (almost surely);
2. $X(\omega)$ has independent and stationary increments;
3. $X(\omega)$ is stochastically continuous;

In some situations we loose the second condition and also call a stochastic process with non-stationary increments as Lévy process. For the beta process example, this corresponds to the case when the concentration function $c(\omega)$ is not a constant, i.e., the inhomogeneous beta process. The beta process is reviewed in Section 2.3.

2.1.2 Pure-jump nondecreasing Lévy process and its underlying Poisson process

By the Lévy-Itô decomposition (Sato, 1999), a Lévy process can be decomposed into a continuous Brownian motion with drift, and a discrete part of a pure-jump process. When a Lévy process $X(\omega)$ only has the discrete part and its jumps are positive, then for $\forall \mathcal{A} \in \mathcal{F}$ the characteristic function of the random variable $X(\mathcal{A})$ is given by:

$$\mathbb{E}\{\mathrm{e}^{juX(\mathcal{A})}\} = \exp\{\int_{\mathbb{R}^+ \times \mathcal{A}} (\mathrm{e}^{jup} - 1)\nu(dp, d\omega)\} \tag{2.1}$$

with ν satisfying the integrability condition (Sato, 1999). The expression in (2.1) defines a category of pure-jump nondecreasing Lévy processes, including most of the Lévy processes currently used in nonparametric Bayesian methods, such as the beta, gamma, Bernoulli, and negative binomial processes. With (2.1), such a Lévy process can be regarded as a Poisson point process on the product space $\mathbb{R}^+ \times \Omega$ with the mean measure ν, called the Lévy measure. On the other hand, if the increments of $X(\omega)$ on any measurable set $\mathcal{A} \in \mathcal{F}$ are regarded as a random measure assigned on the set, then $X(\omega)$ is also a completely random measure. Due to this equivalence, in the following discussion we will not discriminate the pure-jump nondecreasing Lévy process X with its corresponding completely random measure Φ.

2.2 Completely random measure

2.2.1 Definition of completely random measure

A random measure Φ on a measure space $(\Omega, \mathcal{F})$ is termed "completely random" if for any disjoint sets $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_2, \cdots \in \mathcal{F}$ the random variables $\Phi(\mathcal{A}_1), \Phi(\mathcal{A}_2), \Phi(\mathcal{A}_3), \cdots$ are independent. A completely random measure Φ can be split into three independent components:

$$\Phi = \Phi_f + \Phi_d + \Phi_o \tag{2.2}$$

where $\Phi_f = \sum_{\omega \in \mathcal{I}} \phi(\omega)\delta_\omega$ is the fixed component, with the atoms in $\mathcal{I}$ fixed and the *jump* $\phi(\omega)$ random; $\mathcal{I}$ is a countable set in $\mathcal{F}$. The deterministic component Φ_d is a deterministic measure on $(\Omega, \mathcal{F})$. Φ_f and Φ_d are relatively less interesting compared to the third component Φ_o, which is called the ordinary component of Φ. According to (Kingman, 1967), Φ_o is discrete with both random atoms and jumps.

2.2.2 Lévy measure decomposition

In (Kingman, 1967), it is noted that Φ_o can be further split into a countable number of independent parts:

$$\Phi_o = \sum_k \Phi_k, \quad \Phi_k = \sum_{(\phi(\omega),\omega)\in\Pi_k} \phi(\omega)\delta_\omega \tag{2.3}$$

Denote ν as the Lévy measure of (the Lévy process corresponding to) Φ_o, ν_k as the Lévy measure of Φ_k, Π a Poisson process with ν its mean measure, and Π_k a Poisson process with ν_k its mean measure; (2.3) further yields:

$$\nu = \sum_k \nu_k, \quad \Pi = \bigcup_k \Pi_k \tag{2.4}$$

which provides a constructive method for Φ_o: first construct the Poisson process Π_k underlying Φ_k, and then with the superposition theorem (Kingman, 1993) the union of Π_k will be a realization of Φ_o. In Section 3 we show how this general construction method of (2.4) can be applied on pure-jump nondecreasing Lévy processes of increasing interest in machine learning, with an emphasis on the beta and gamma processes, and their generalizations. And before that we will review the beta process and gamma process.

2.3 Beta process

A beta process was first proposed by (Hjort, 1990) in survival analysis. Beta process is a Lévy process with beta-distributed increments. $B \sim \mathrm{BP}(c(\omega), \mu)$ is a beta process if

$$B(d\omega) \sim \mathrm{Beta}(c(\omega)\mu(d\omega), c(\omega)(1 - \mu(d\omega))) \tag{2.5}$$

where μ is the base measure on measure space $(\Omega, \mathcal{F})$ and a positive function $c(\omega)$ the concentration function. Expression (2.5) indicates that the increments of the beta process are independent, which makes it a special case of the Lévy process family. The Lévy measure of the beta process is

$$\nu(d\pi, d\omega) = c(\omega)\pi^{-1}(1 - \pi)^{c(\omega)-1} d\pi \mu(d\omega) \tag{2.6}$$

where $\mathrm{Beta}(0, c(\omega)) = c(\omega)\pi^{-1}(1 - \pi)^{c(\omega)-1}$ is an *improper* beta distribution since its integral over $(0, 1)$ is infinite. As a result, its *underlying Poisson process*, *i.e.*, the Poisson process with ν as its mean measure on the product space $\Omega \times (0, 1)$, denoted Π, has an infinite number of points drawn from ν, yielding

$$B = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i} \tag{2.7}$$

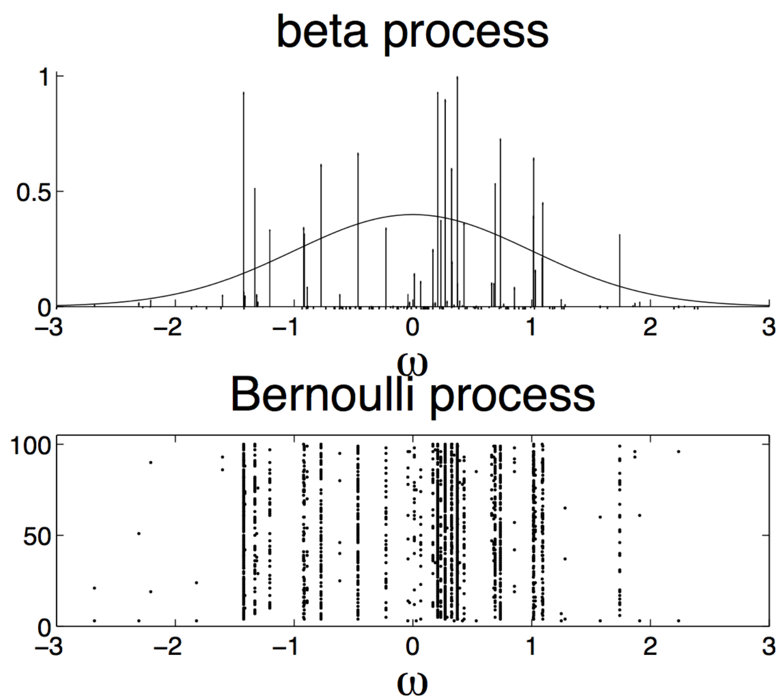


FIGURE 2.1: Beta process: **Top row:** beta process with a Gaussian base measure. **Bottom row:** 100 independent Bernoulli processes with the beta process as the prior.

where π_i is the jump (increment) which happens at the atom ω_i. Real variable $\gamma = \mu(\Omega)$ is termed the mass parameter of B, and we assume $\gamma < \infty$.

2.4 Gamma process

A gamma process (Applebaum, 2009) is a Lévy process with independent gamma increments. The gamma process is traditionally parameterized with a shape measure and a scale function: $G \sim \Gamma\mathrm{P}(\alpha, \theta(\omega))$ where α is the shape measure on a measure space $(\Omega, \mathcal{F})$, and the scale $\theta(\omega)$ a positive function. A gamma process can be intuitively defined by its increments on infinitesimal sets:

$$G(d\omega) \sim \mathrm{Gamma}(\alpha(d\omega), \theta(\omega)) \tag{2.8}$$

When $\theta(\omega) = \theta$ is a scalar, the gamma process is called homogeneous. The gamma process can also be expressed in the form with a base measure G_0 and a concentration $c(\omega)$, with $c = 1/\theta$ and $G_0 = \theta\alpha$ (Jordan, 2009), to conform with other stochastic processes widely used in machine learning, such as the Dirichlet process. However, the discussion in the Lévy measure decomposition discussed in Chapter 3 will stick to the traditional form given by (2.8).

As a pure-jump Lévy process, the gamma process can be regarded as a Poisson process on the product space $\Omega \times \mathbb{R}^+$ with mean measure ν:

$$\nu(dp, d\omega) = p^{-1}\mathrm{e}^{-\frac{p}{\theta(\omega)}} dp\alpha(d\omega) \tag{2.9}$$

where $\mathrm{Gamma}(0, \theta(\omega)) = p^{-1}\mathrm{e}^{-\frac{p}{\theta(\omega)}}$ is an improper gamma distribution with an infinite integral on $\mathbb{R}^+$, which yields the expression of G:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\omega_i} \tag{2.10}$$

3

Lévy Measure Decompositions for the Beta and Gamma Processes

In this chapter we develop explicit and simple decompositions by following the conjugacy principle for two widely used Lévy processes, the beta and gamma processes. The conjugacy means that the decompositions are manifested by leveraging the forms of conjugate likelihoods to the Lévy measures. The decompositions bring new perspectives on the beta and gamma processes, with associated properties analyzed here in detail. The decompositions are constituted in terms of an infinite set of sub-processes of form convenient for computation. Since the number of sub-processes is infinite, a truncation analysis is also presented, of interest for practical use. We show some posterior properties of such decompositions, with the beta process as an example. We also extend the decomposition to the *symmetric* gamma process (positive and negative jumps), suggesting that the Lévy measure decomposition is applicable for other pure-jump Lévy processes represented by their Lévy measures. Summarizing the main contributions of our methods:

- We constitute Lévy measure decompositions for the beta, stable-beta, gamma, generalized gamma and symmetric gamma processes via the principle of conjugacy, providing new perspectives on these processes.

- The decomposition of the beta process unifies the constructions in (Thibaux and Jordan, 2007a), (Teh and Görür, 2009), and (with a different decomposing method) (Paisley et al., 2010), and a new generative construction for the gamma process and its variations is derived.

- Truncation analyses and posterior properties for such decompositions are presented for practical use.

3.1 Lévy measure decomposition for the beta process

3.1.1 Lévy measure decomposition

The infinite integral of the improper beta distribution inspires a decomposition of the improper distribution with an infinite number of *proper* distributions. The singularity in the improper beta distribution is manifested from π^{-1}. Since $\pi \in (0,1)$, the geometric series expansion yields

$$\pi^{-1} = \sum_{k=0}^{\infty}(1-\pi)^k, \quad \pi \in (0,1) \tag{3.1}$$

and substituting (3.1) in (2.6), with manipulation detailed in the Appendix, we have the Lévy measure decomposition theorem of the beta process:

Theorem 1 *For a beta process $B \sim \mathrm{BP}(c(\omega),\mu)$ with base measure μ and concentration $c(\omega)$, denote Π as its underlying Poisson process and ν the* Lévy *measure, then B and Π can be expressed as*

$$\Pi = \bigcup_{k=0}^{\infty}\Pi_k\ , \quad B = \sum_{k=0}^{\infty} B_k \tag{3.2}$$

where B_k *is a Lévy process with* Π_k *its underlying Poisson process. The Lévy measure* ν_k of B_k *is a decomposition of* ν:

$$\begin{aligned}
&\nu = \sum_{k=0}^{\infty} \nu_k \\
&\nu_k(d\pi, d\omega) = \mathrm{Beta}(1, c(\omega) + k) d\pi \mu_k(d\omega) \\
&\mu_k(d\omega) = \frac{c(\omega)}{c(\omega) + k} \mu(d\omega)
\end{aligned} \tag{3.3}$$

where $\mathrm{Beta}(1, c(\omega)+k)$ *is the PDF of beta distribution with parameters* 1 *and* $c(\omega)+k$.

Theorem 1 is the beta process instantiation of the completely random measure decomposing in (2.4), which indicates that the underlying Poisson process Π of the beta process B is the superposition of an infinite number of independent Poisson processes $\{\Pi_k\}_{k=0}^{\infty}$, with ν_k the mean measure of Π_k and μ_k the mean measure of the restriction of Π_k on Ω. As a result, the beta process B can be expressed as a sum of an infinite number of independent Lévy processes $\{B_k\}_{k=0}^{\infty}$ with $\{\Pi_k\}_{k=0}^{\infty}$ the underlying Poisson process. The independence of $\{\Pi_k\}_{k=0}^{\infty}$ and $\{B_k\}_{k=0}^{\infty}$ w.r.t. index k is justified by the fact that both μ and $c(\omega)$ are fixed parameters.

3.1.2 The Lévy process B_k

It is interesting to study the properties of B_k, such as the expectation and variance. Denoting $\mathcal{B}_k(d\omega) = \frac{1}{c(\omega)+k+1}\mu_k(d\omega)$ as the base measure of B_k, for $\forall \mathcal{A} \in \mathcal{F}$:

$$\begin{aligned}
&\mathbb{E}(B_k(\mathcal{A})) = \int_{\mathcal{A}} \mathcal{B}_k(d\omega) = \mathcal{B}_k(\mathcal{A}) \\
&\mathrm{Var}(B_k(\mathcal{A})) = \int_{\mathcal{A}} \frac{2}{c(\omega) + k + 2} \mathcal{B}_k(d\omega)
\end{aligned} \tag{3.4}$$

It is noteworthy that the Lévy process B_k is no longer a beta process, since (2.5) is not satisfied. By Theorem 1, the jumps of B_k follow a *proper* beta distribution parameterized by the concentration function $c(\omega)$ and the index k, and μ_k determines the locations where the jumps happen. Since $\{B_k\}_{k=0}^{\infty}$ are independent w.r.t. the index k, with Theorem 1:

$$\begin{aligned}\sum_{k=0}^{\infty} \mathbb{E}(B_k(\mathcal{A})) &= \mathbb{E}(B(\mathcal{A})) \\ \sum_{k=0}^{\infty} \mathrm{Var}(B_k(\mathcal{A})) &= \mathrm{Var}(B(\mathcal{A}))\end{aligned} \tag{3.5}$$

The detailed procedure to derive (3.4) and (3.5) is given in the Appendix.

3.1.3 Simulating the beta process

Poisson superposition simulation

Theorem 1 reveals that the underlying Poisson process of a beta process is a superposition of an infinite number of Poisson processes, each of which has a *finite* set of atoms. This perspective also provides a simulation procedure for the beta process: first, the Poisson process Π_k is sampled for all $k = 0, 1, 2, \cdots$, (here we term the index k as the "round" of the simulation); then take the union of the samples of each Π_k as a realization of the Poisson process Π. With the marking theorem (Kingman, 1993) implicitly applied, the simulation procedure of the beta process is as follows:

Simulation procedure: For round k:

1: Sample the number of points for Π_k: $n_k \sim \mathrm{Poisson}(\int_\Omega \mu_k(d\omega))$;

2: Sample n_k points from μ_k: $\omega_{ki} \overset{\text{i.i.d.}}{\sim} \frac{\mu_k}{\int_\Omega \mu_k(d\omega)}$, for $i = 1, 2, \cdots, n_k$;

3: Sample $B_k(\omega_{ki}) \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, c(\omega_{ki}) + k)$, for $i = 1, 2, \cdots, n_k$;

Then the union $\bigcup_{k=0}^{\infty}\{(\omega_{ki}, B_k(\omega_{ki}))\}_{i=1}^{n_k}$ is a realization of Π (and equivalently of B).

We refer to the above simulation procedure as the *Poisson superposition simulation*, for the central role of the Poisson superposition. The especially convenient case is when the beta process is homogeneous, *i.e.*, $c(\omega) = c$ is a constant. In this case $\{\omega_{ki}\}_{i=1}^{n_k}$ for all rounds k are drawn from the same distribution μ/γ; and n_k is drawn from Poisson$(\frac{c\gamma}{c+k})$. For round k, both the number of points and the jumps statistically diminish as k increases, suggesting that the infinite sum in (3.2) may be truncated as $B = \sum_{k=0}^{K} B_k$ for large K, with minimal impact. Such truncation effects are investigated in detail in Section 3.1.4.

Related work

In (Thibaux and Jordan, 2007a) the authors derived the above simulation procedure for the homogeneous case within the beta-Bernoulli process context, which is shown here a necessary result of the Lévy measure decomposition. The same decomposing manipulation of Theorem 1 can be also applied to the stable beta process (Teh and Görür, 2009) which yields:

$$\nu_k = \mathrm{Beta}(1-\sigma, c(\omega)+\sigma+k)d\pi \cdot \frac{\Gamma(c(\omega)+\sigma+k)\Gamma(c(\omega)+1)}{\Gamma(c(\omega)+k+1)\Gamma(c(\omega)+\sigma)}\mu(d\omega) \tag{3.6}$$

It is noteworthy that the decomposition procedure described in Theorem 1 is not the only Lévy measure decomposing method for the beta process. The work of (Paisley et al., 2012) and (Broderick et al., 2011) show that the stick-breaking construction of the beta process in (Paisley et al., 2010) is indeed a result of another way of decomposing the Lévy measure of the beta process. We next analyze the truncation property of the construction described in Section 3.1.3 and make comparison with the construction of beta process in (Paisley et al., 2010).

3.1.4 Truncation analysis

Since the Poisson superposition simulation operates in rounds, it is natural to analyze the distance between the true beta process B and its truncation $\sum_{k=0}^{K} B_k$, with truncation at round K. A metric for such distance is the $\mathcal{L}_1$ norm:

$$||B - \sum_{k=0}^{K} B_k||_1 = \mathbb{E}|B - \sum_{k=0}^{K} B_k| = \int_\Omega \frac{\mu_{K+1}(d\omega)}{\gamma} \tag{3.7}$$

The expectation in (3.7) is w.r.t. the normalized measure ν/γ, which yields $\|B\|_1 = 1$. When B is homogeneous, (3.7) reduces to $\frac{c}{c+K+1}$, which indicates that the $\mathcal{L}_1$ distance decreases at a rate of $\mathcal{O}(\frac{1}{K})$. For the stick-breaking construction of beta process described in (Paisley et al., 2010), the $\mathcal{L}_1$ distance is: $(\frac{c}{c+1})^{K+1}$.

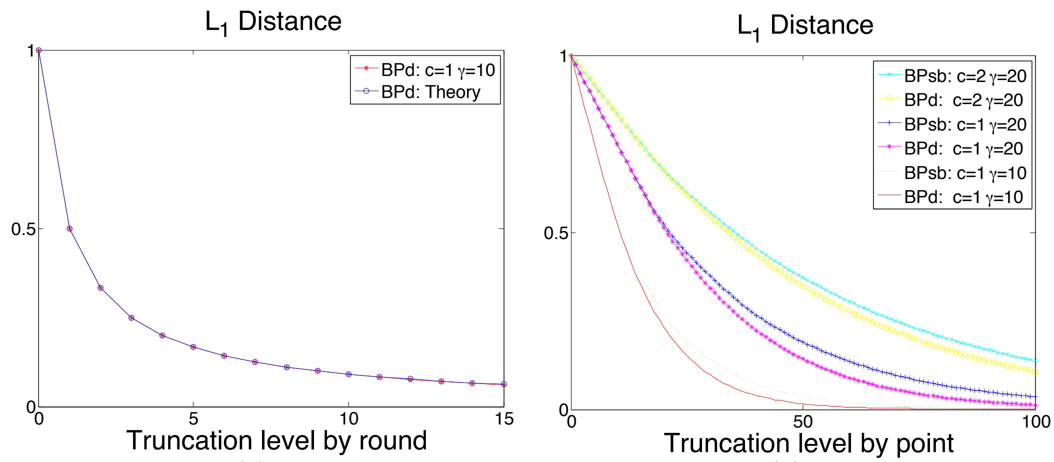
Another metric is the $\mathcal{L}_1$ distance between the marginal likelihood of a set of data $\boldsymbol{b} = b_{1:M}$, with $m_\infty(\boldsymbol{b})$ denotes the marginal likelihood (here the likelihood is a Bernoulli process) with prior B, and $m_K(\boldsymbol{b})$ for $\sum_{k=0}^{K} B_k$. This metric was applied on the truncated Indian buffet process (Doshi et al., 2009) and truncated stick-breaking construction of the beta process (Paisley et al., 2012), which indicates

$$\frac{1}{4}\int |m_\infty(\boldsymbol{b}) - m_K(\boldsymbol{b})| d\boldsymbol{b} \leqslant \Pr(\exists k > K,\ 1 \leqslant i \leqslant n_k,\ 1 \leqslant m \leqslant M,\ \text{s.t. } b_{ki}^m = 1) \tag{3.8}$$

where $b_{1:M} \overset{\text{i.i.d.}}{\sim} \text{BeP}(B)$ are drawn from a Bernoulli process with base measure B; $b_{ki}^m = b_m(\omega_{ki})$ is the m^{th} realization of the Bernoulli process at atom ω_{ki}. For the truncation $\sum_{k=0}^{K} B_k$ it can be shown that the RHS of (3.8) is bounded by:

$$\text{RHS of (3.8)} \leqslant 1 - \exp(-M \int_\Omega \mu_{K+1}(d\omega)) \tag{3.9}$$

For the homogeneous case, the bound of (3.9) is $1 - \exp(-M\gamma\frac{c}{c+K+1})$. For the stick-breaking construction of beta process, the bound is given by: $1 - \exp(-M\gamma(\frac{c}{c+1})^{K+1})$



(a) (b)

FIGURE 3.1: Simulation of the truncation errors of the beta process decomposition presented in Theorem 1. (a) Comparison of the truncation errors by round yielded by simulating the beta process decomposition with the theoretical analysis. (b) Comparison of the truncation errors by point yielded by simulating the beta process decomposition and stick-breaking beta process, with different c and γ.

(Paisley et al., 2012).

In order to analyze the bound w.r.t. the truncation level by number of atoms, denote $I_K = \sum_{k=0}^{K} n_k$ as the total number of atoms in $\sum_{k=0}^{K} B_k$. Since $K \sim \mathcal{O}(\mathrm{e}^{\frac{\mathbb{E}(I_K)}{c\gamma}})$, it is proved that (3.7) and the bound in (3.9) decreases at a faster rate w.r.t. I than the stick-breaking construction of beta process. This indicates that the simulation procedure described in Section 3.1.3 follows a steeper statistically-decreasing order. The proof is presented in the Appendix.

Figure 3.1 shows the simulation results of the truncation error analysis. Figure 3.1(a) compares the simulated truncation errors by round of the beta process decomposition (denoted as BPd) for a beta process with $c = 1, \gamma = 10$ with the theoretical results given by (3.7). In (a) the simulated results accord with the theoretical analysis. Figure 3.1(b) compares the truncation errors by point of the beta process decomposition (denoted as BPd) for beta processes with different c and γ, with the stick-breaking beta process (denoted as BPsb). It is shown in (b) that the truncation errors of the beta process decomposition follow steeper decreases than the stick-breaking beta process, as we theoretically proved.

However, it is noteworthy that by the first step of the simulation procedure in Section 3.1.3, n_k is statistically decreasing with the round k. When k goes large, it is typical that it takes many times to draw a non-zero n_k by following the first step of the simulation procedure. While in the stick-breaking construction of the beta process (Paisley et al., 2012), n_k is statistically unchanged with the round k, which is more efficient in simulating the beta process, although with higher truncation errors.

3.1.5 Posterior estimation

The goal of the inference is to estimate the beta process B from a set of observed data $\boldsymbol{b}$ with prior $\mathrm{BP}(c, \mu)$. The data $\boldsymbol{b} = b_{1:M}$ is the same as in Section 3.1.4, which can be expressed as:

$$b_m = \sum_{i=1}^{\infty} b_{i,m} \delta_{\omega_i}, \quad m = 1, 2, \cdots, M \tag{3.10}$$

where each $b_{i,m} \in \{0, 1\}$.

Posterior of B_k

Since $B|\boldsymbol{b} \sim \mathrm{BP}(c + M, \frac{c\mu}{c+M} + \frac{\sum_{m=1}^{M} b_m}{c+M})$ (Thibaux and Jordan, 2007a), the base measure of $B|\boldsymbol{b}$ is a measure with positive masses assigned on single atoms. Theorem 1 is still applicable to this beta process with mixed type of base measure, which yields

$$
\begin{aligned}
B' &= \sum_{k=0}^{\infty} B'_k \\
\nu'_k &= \mathrm{Beta}(1, c + M + k)\mu'_k \\
\mu'_k &= \frac{c\mu}{c + M + k} + \frac{\sum_{m=1}^{M} b_m}{c + M + k}
\end{aligned} \tag{3.11}
$$

where the B', B'_k, ν'_k, and μ'_k are the posterior counterparts of B, B_k, ν_k, and μ_k.

Posterior estimation of π_i:

Since each μ_k has a mass $\frac{\sum_{m=1}^{M} b_{i,m}}{c+M+k}$ at the atom ω_i, each B_k will contribute Poisson($\frac{\sum_{m=1}^{M} b_{i,m}}{c+M+k}$) draws with the jumps following the distribution $\mathrm{Beta}(1, c+M+k)$ at the atom ω_i, whose sum is the π_i. Thus the posterior estimation of π_i is given by

$$
\begin{aligned}
\pi_i|\boldsymbol{b} &= \sum_{k=0}^{\infty} \sum_{h=1}^{H_k} b_{kh} \\
H_k &\sim \mathrm{Poisson}(\frac{\sum_{m=1}^{M} b_{i,m}}{c + M + k}) \\
b_{kh} &\sim \mathrm{Beta}(1, c + M + k)
\end{aligned} \tag{3.12}
$$

from which it can be verified that $\mathbb{E}(\pi_i|\boldsymbol{b}) = \frac{\sum_{m=1}^{M} b_{i,m}}{c+M}$, the same as the posterior of π_i without decomposition: $\mathrm{Beta}(\sum_{m=1}^{M} b_{i,m}, c + M - \sum_{m=1}^{M} b_{i,m})$.

For the π_i with no observations, *i.e.*, $\sum_{m=1}^{M} b_{i,m} = 0$, only a particular B_k will contribute to π_i. In this case, first the round k to which π_i belongs is drawn, then π_i is drawn from the beta distribution of that round:

$$
\begin{aligned}
\pi_i &\sim \mathrm{Beta}(1, c + M + k) \\
k &\sim \mathrm{MP}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} \propto \sum_{k=0}^{\infty} \frac{1}{c + M + k} \delta_k
\end{aligned} \tag{3.13}
$$

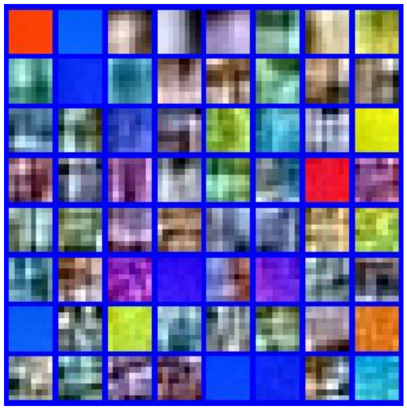
where MP($\boldsymbol{\alpha}$) is a multinomial process with probability vector $\boldsymbol{\alpha}$, and $\boldsymbol{\alpha}$ is proportional to the average number of points in each round. Since in practical processing $\boldsymbol{\alpha}$ is always to be truncated with a truncation level K, by the analysis in Section 3.1.4, (3.13) provides a way to estimate the π_i within the first K rounds. And π_i in each round are of statistically different importance, contrasted to the evenly assigned mass in the Indian buffet process.

3.1.6 Feature learning experiment with Lévy measure decomposition of beta process

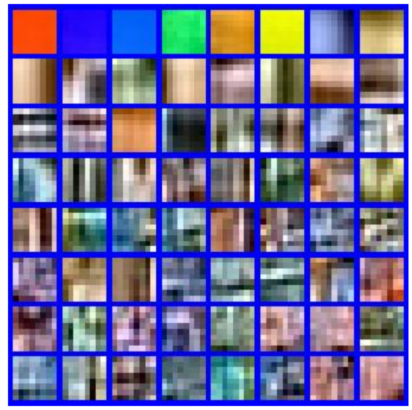
We apply the Lévy measure decomposition for beta process described in Theorem 1 in the image inpainting problem considered in (Zhou et al., 2009), based upon a beta process factor analysis model (Paisley and Carin, 2009). In experiments we performed with such a model, using a Gibbs sampler, the beta process prior was implemented using the procedure discussed in Section 3.1.3, with the posterior estimation in Section 3.1.5 applied for inference. As shown in Figure 3.2, the proposed representation infers a dictionary with the "important" dictionary elements captured by the low-index members (see the discussion in Section 3.1.3). The model prioritized the first three dictionary elements as being pure colors, specifically red, green, and blue, with the important structured dictionary elements following (and no other pure-color dictionary elements, while in (Zhou et al., 2009) many – seemingly redundant – pure-color dictionary elements are inferred). This "clean" inference of prioritized dictionary elements may be responsible for our also higher observed PSNR in signal recovery, compared to the result given in (Zhou et al., 2009).

3.1.7 Relating the IBP and beta process

The study of the beta process through its Lévy measure, as discussed here, also uncovers a connection between the Indian buffet process (IBP) (Griffiths and Ghahramani, 2005) and the beta process, by their Lévy measures. The IBP with prior



(a)



(b)

FIGURE 3.2: Comparison of the feature dictionary elements learned in image inpainting. (a) Features learned via IBP prior. (b) Features learned via the beta process decomposition.

$\pi_i \sim \text{Beta}(c\frac{\gamma}{N}, c)$ can be regarded as a Lévy process with the Lévy measure given as:

$$\nu_{\text{IBP}} = \frac{N}{\gamma}\text{Beta}(c\frac{\gamma}{N}, c)d\pi\mu(d\omega) \tag{3.14}$$

here N is the same as the K in (Griffiths and Ghahramani, 2005). It can be proved that:

$$\nu_{\text{IBP}} \overset{N\to\infty}{=} \nu \tag{3.15}$$

which indicates that the beta process is the limit of the IBP with $N \to \infty$. The detailed proof of (3.15) is presented in the Appendix. Thus the IBP is like a "mosaic" approximation of beta process, which becomes finer with N increases.

3.2 Lévy measure decomposition for gamma process

3.2.1 Lévy measure decomposition

Like the beta process, the Lévy measure of the gamma process is characterized by an improper distribution. However, unlike the beta process, the decomposition of the Lévy measure of the gamma process comes from the exponential part. With the details shown in the Appendix, the gamma process G can be decomposed into two parts:

$$G = \Gamma_1 + \Gamma\mathrm{P}(\alpha, \theta(\omega)/2) \tag{3.16}$$

The second term in (3.16) is a gamma process with the same shape measure, and half the scale of the gamma process G; the first term Γ_1 is a Lévy process with the Lévy measure $\sum_{h=1}^{\infty} \mathrm{Gamma}(h, \frac{\theta(\omega)}{2}) dp \frac{\alpha(d\omega)}{2^h h}$. Here $\mathrm{Gamma}(h, \frac{\theta(\omega)}{2})$ is the PDF of the gamma distribution, with shape parameter h and scale parameter $\frac{\theta(\omega)}{2}$.

Further decomposing the exponential part of the gamma process $\Gamma\mathrm{P}(\alpha, \theta(\omega)/2)$ in (3.16) yields $G = \Gamma_1 + \Gamma_2 + \Gamma\mathrm{P}(\alpha, \theta(\omega)/3)$, bearing a gamma process with the same shape and with the scale parameter further decreased. Repeating this manipulation, we obtain the Theorem 2:

Theorem 2 *A gamma process $G \sim \Gamma\mathrm{P}(\alpha, \theta(\omega))$ with shape measure α and scale $\theta(\omega)$ can be decomposed as:*

$$\begin{aligned} G &= \sum_{k=1}^{\infty} \Gamma_k, \ \Gamma_k = \sum_{h=1}^{\infty} \Gamma_{kh}, \ \nu_k = \sum_{h=1}^{\infty} \nu_{kh} \\ \nu_{kh} &= \mathrm{Gamma}(h, \frac{\theta(\omega)}{k+1}) dp \frac{\alpha(d\omega)}{(k+1)^h h} \end{aligned} \tag{3.17}$$

with Γ_k, Γ_{kh} Lévy processes with ν_k, ν_{kh} their Lévy measures.

Theorem 2 is the gamma process instantiation of (2.4), which indicates that G can be expressed as the sum of an infinite number of Lévy processes $\Gamma_k, k = 1, 2, \cdots$, where Γ_k is also the sum of an infinite number of Lévy processes $\Gamma_{kh}, h = 1, 2, \cdots$.

3.2.2 *Lévy processes Γ_k and Γ_{kh}*

In order to obtain further insights into the gamma process G in Theorem 2, the expectations and variances of Γ_k and Γ_{kh} on any measurable set $\mathcal{A} \in \mathcal{F}$ are given:

$$\begin{aligned} \mathbb{E}(\Gamma_{kh}(\mathcal{A})) &= \frac{\int_{\mathcal{A}} \theta(\omega)\alpha(d\omega)}{(k+1)^{h+1}} \\ \mathbb{E}(\Gamma_k(\mathcal{A})) &= \frac{\int_{\mathcal{A}} \theta(\omega)\alpha(d\omega)}{k(k+1)} \end{aligned} \tag{3.18}$$

For the variances of Γ_k and Γ_{kh}:

$$\begin{aligned} \mathrm{Var}(\Gamma_{kh}(\mathcal{A})) &= \frac{(h+1)}{(k+1)^{h+2}} \int_{\mathcal{A}} \theta^2(\omega)\alpha(d\omega) \\ \mathrm{Var}(\Gamma_k(\mathcal{A})) &= [\frac{1}{k^2} - \frac{1}{(k+1)^2}] \int_{\mathcal{A}} \theta^2(\omega)\alpha(d\omega) \end{aligned} \tag{3.19}$$

Since the Lévy processes Γ_k are independent w.r.t. k, with analogy to (3.5) it can be verified that the expectation and variance of Γ_k sum to the expectation of variance of G. The derivations in this section are presented in the Appendix.

3.2.3 *Simulation of gamma process*

Parallel to the simulation of beta process in Section 3.1.3, a simulation procedure of the gamma process is presented:

Simulation procedure: Sample the Lévy process Γ_{kh}:

1: Sample the number of points for Γ_{kh}: $n_{kh} \sim \mathrm{Poisson}(\gamma/(k+1)^h h)$;

2: Sample n_{kh} points from α: $\omega_{khi} \overset{\text{i.i.d.}}{\sim} \frac{\alpha}{\gamma}$, for $i = 1, 2, \cdots, n_{kh}$;

3: Sample $\Gamma_{kh}(\omega_{khi}) \overset{\text{i.i.d.}}{\sim} \mathrm{Gamma}(h, \frac{\theta(\omega_{khi})}{k+1})$, for $i = 1, 2, \cdots, n_{kh}$;

where $\gamma = \int_\Omega \alpha(d\omega)$ is the mass of the shape measure. Then the union $\bigcup_{k=1}^{\infty} \bigcup_{h=1}^{\infty} (\omega_{khi}, \Gamma_{kh}(\omega_{khi}))_{i=1}^{n_{kh}}$ is a realization of the gamma process G. An advantage of the above simulation procedure compared to the simulation procedure of the beta process in Section 3.1.3 is that independent of whether the gamma process is homogeneous or inhomogeneous, ω_{khi} is always drawn from a fixed distribution α/γ. Like with the beta process construction in Section 3.1.3, for the gamma process simulation procedure, as k increases the expected number of new points and the expected jumps decrease, again suggesting accurate truncation.

3.2.4 Truncation analysis

Since in the simulation procedure in Section 3.2.3 the index k and h both go to infinity, it is practical to analyze the distance between the true gamma process and the truncated one. To measure such a distance, we apply the $\mathcal{L}_1$ norm described in Section 3.1.4:

$$||G - \sum_{k=1}^{K} \sum_{h=1}^{H} \Gamma_{kh}||_1 = \mathbb{E}|G - \sum_{k=1}^{K} \sum_{h=1}^{H} \Gamma_{kh}| \tag{3.20}$$

where the expectation in (3.20) is w.r.t. the normalized measure $\nu / \int_\Omega \theta(\omega)\alpha(d\omega)$ with $||G||_1 = 1$; and K and H are the truncation level of k and h. Then for the situation with $H = \infty$:

$$\|G - \sum_{k=1}^{K} \sum_{h=1}^{\infty} \Gamma_{kh}\|_1 = \frac{1}{K+1} \tag{3.21}$$

which indicates a $\mathcal{O}(\frac{1}{K})$ decreasing rate as same as the truncated beta process shown in (3.7). It is noteworthy that Γ_1 alone accounts for on average half the mass of G. When H is finite, a remaining distance $\sum_{k=1}^{K} \frac{1}{k(k+1)^{H+1}}$ is added.

3.2.5 Posterior estimation

In this section the posterior estimation of the gamma process decomposition proposed in Theorem 2 is presented. For convenience, here a gamma process is represented with its shape measure $\alpha(\omega)$ and rate function $\beta(\omega)$, where $\beta(\omega) = \frac{1}{\theta(\omega)}$.

Besides, to present the posterior analysis of the gamma process decomposition in a parallel form to the beta process case given in Section 3.1.5, we assume the likelihood is yielded with M independent Poisson process with the prior gamma process as the base measure $G(d\omega) \sim \Gamma\mathrm{P}(\alpha(d\omega), \beta(\omega))$. The observed data $\boldsymbol{n} = n_{1:M}$, which can be expressed as:

$$n_m = \sum_{i=1}^{\infty} n_{i,m}\delta_{\omega_i} \tag{3.22}$$

Given the gamma process expressed as $G = \sum_{i=1}^{\infty} p_i\delta_{\omega_i}$, the likelihood is given by

$$\boldsymbol{n}|G = \prod_{m=1}^{M}\prod_{i=1}^{\infty} \frac{e^{-p_i}{p_i}^{n_{i,m}}}{n_{i,m}!} \tag{3.23}$$

Then the posterior is given by

$$G' \sim \Gamma P(\alpha + \sum_{m=1}^{M} n_m, \beta + M) \tag{3.24}$$

And the posterior estimation for the gamma process decomposition as proposed in Theorem 2 is to decompose the gamma process with the shape measure $\alpha + \sum_{m=1}^{M} n_m$ and rate function $\beta + M$, analogous to the posterior estimation of the beta process decomposition as discussed in Section 3.1.5, with ν'_{kh} given by

$$\nu'_{kh} = \mathrm{Gamma}(h, \frac{1}{(\beta(\omega) + M)(k+1)})dp\frac{\alpha(d\omega) + \sum_{m=1}^{M} n_m}{(k+1)^h h} \tag{3.25}$$

3.2.6 Generalized gamma process and symmetric gamma process

Theorem 2 can be easily extended to some variations of the gamma process. Here we give the examples of the generalized gamma process (Brix, 1999) and symmetric gamma process (Çinlar, 2010).

The generalized gamma process extends the ordinary gamma process by adding a parameter $0 < \sigma < 1$, whose Lévy measure is $\frac{1}{\Gamma(1-\sigma)} p^{-\sigma-1} \mathrm{e}^{-\frac{p}{\theta(\omega)}} dp\alpha(d\omega)$. Then with the same decomposition procedure, it is straightforward that the Lévy measure for Γ_{kh} of the generalized gamma process will change to $\nu_{kh} = \mathrm{Gamma}(h - \sigma, \frac{\theta(\omega)}{k+1}) dp \frac{\alpha(d\omega)}{\Gamma(1-\sigma)(k+1)^h h}$.

The symmetric gamma process is a Lévy process whose increments are the differences of two gamma-distributed variables with the same law, whose Lévy measure is $|p|^{-1} \mathrm{e}^{-\frac{|p|}{\theta(\omega)}} dp\alpha(d\omega)$. Since there can be negative increments, the symmetric gamma process is not a completely random measure. However, the same decomposition procedure is still applicable, yielding $\nu_{kh} = \mathrm{Gamma}(|p| \big| h, \frac{\theta(\omega)}{k+1}) dp \frac{2\alpha(d\omega)}{(k+1)^h h}$, where the distribution $\mathrm{Gamma}(|p| \big| h, \frac{\theta(\omega)}{k+1})$ is to first draw $|p|$ from $\mathrm{Gamma}(h, \frac{\theta(\omega)}{k+1})$, then decide the sign of p through a symmetric Bernoulli distribution.

4

Kernel Beta Process

The beta process as discussed in Section 2.3 and 3.1 is an example of Lévy processes (Kingman, 2002). And in Section 3.1.7 we proved that IBP can be regarded as the limit of the beta process. The beta-Bernoulli constructions have found significant utility in factor analysis (Knowles and Ghahramani, 2007; Zhou et al., 2009), in which one wishes to infer the number of factors needed to represent data of interest. Here we refer to the base measures used in the feature models as the "feature-learning measures."

Another example of Lévy processes is the gamma process as discussed in Section 2.4 and 3.2; the *normalized* gamma process is well known as the Dirichlet process (Ferguson, 1973; Sethuraman, 1994b) or the Chinese restaurant process (Pitman, 1995), and is widely used as the base measure in the mixture models. A key characteristic of such models with the beta process and Dirichlet process is that the data samples are assumed exchangeable, meaning that the order/indices of the data may be permuted with no change in the model.

An important line of research concerns removal of the assumption of exchangeability, allowing incorporation of covariates (*e.g.*, spatial/temporal coordinates that may be available with the data). As an example, MacEachern introduced the *dependent* Dirichlet process (MacEachern, 1999). In the context of feature learning, the phylogenetic IBP removes the assumption of sample exchangeability by imposing prior knowledge on inter-sample relationships via a tree structure (Miller et al., 2008). The form of the tree may be constituted as a result of covariates that are available with the samples, but the tree is not necessarily unique. A dependent IBP (dIBP) model has been introduced recently, with a hierarchical Gaussian process (GP) used to account for covariate dependence (Williamson et al., 2010); however, the use of a GP may constitute challenges for large-scale problems. Recently a dependent hierarchical beta process (dHBP) has been developed, yielding encouraging results (Zhou et al., 2011). However, the dHBP has the disadvantage of assigning a kernel to each data sample, and therefore it scales unfavorably as the number of samples increases.

In this section we develop a new Lévy process prior, termed the kernel beta process (KBP), which yields an uncountable number of *covariate-dependent* feature-learning measures, with the beta process a special case. This model may be interpreted as inferring covariates $\boldsymbol{x}_i^*$ for each feature (dish), indexed by i. The generative process by which the nth data sample, with covariates $\boldsymbol{x}_n$, selects features may be viewed as a two-step process. First the nth customer (data sample) decides whether to "examine" dish i by drawing $z_{ni}^{(1)} \sim \text{Bernoulli}(K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*))$, where ψ_i^* are dish-dependent kernel parameters that are also inferred (the $\{\psi_i^*\}$ defining the meaning of proximity/locality in covariate space). The kernels are designed to satisfy $K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*) \in (0, 1]$, $K(\boldsymbol{x}_i^*, \boldsymbol{x}_i^*; \psi_i^*) = 1$, and $K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*) \to 0$ as $\|\boldsymbol{x}_n - \boldsymbol{x}_i^*\|_2 \to \infty$. In the second step, if $z_{ni}^{(1)} = 1$, customer n draws $z_{ni}^{(2)} \sim \text{Bernoulli}(\pi_i)$, and if $z_{ni}^{(2)} = 1$, the feature associated with dish i is employed by data sample n. The parameters $\{\boldsymbol{x}_i^*, \psi_i^*, \pi_i\}$ are inferred by the model. After computing the posterior distribution on model parameters, the number of kernels required to represent the measures is defined by the number of features employed from the buffet (typically small relative to the data size); this is a significant computational savings relative to (Zhou et al., 2011; Williamson et al., 2010), for which the complexity of the model is tied to the number of data samples, even if a small number of features are ultimately employed.

In addition to introducing this new Lévy process, we examine its properties, and demonstrate how it may be efficiently applied in important data analysis problems. The hierarchical construction of the KBP is fully conjugate, admitting convenient Gibbs-sampling (complicated sampling methods were required for the method in (Zhou et al., 2011)). To demonstrate the utility of the model we consider image-processing and music-analysis applications, for which state-of-the-art performance is demonstrated compared to other relevant methods.

4.1 Kernel Beta Process

4.1.1 Review of beta and Bernoulli processes

For a beta process $B \sim \text{BP}(c, B_0)$, where $c(\omega)$ is the concentration function and B_0 the base measure, with the discuss in Section 2.3 the Lévy measure of $\text{BP}(c, B_0)$ is given by

$$\nu(d\pi, d\omega) = c(\omega)\pi^{-1}(1-\pi)^{c(\omega)-1} d\pi B_0(d\omega) \tag{4.1}$$

To draw B, one draws a set of points $(\omega_i, \pi_i) \in \Omega \times [0, 1]$ from a Poisson process with measure ν, yielding

$$B = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i} \tag{4.2}$$

where δ_{ω_i} is a unit point measure at ω_i; B is therefore a discrete measure, with probability one. The infinite sum in (4.2) is a consequence of drawing Poisson(λ) atoms $\{\omega_i, \pi_i\}$, with $\lambda = \int_\Omega \int_{[0,1]} \nu(d\omega, d\pi) = \infty$. Additionally, for any set $\mathcal{A} \subset \mathcal{F}$,

$B(\mathcal{A}) = \sum_{i:\ \omega_i \in \mathcal{A}} \pi_i$.

If $Z_n \sim \mathrm{BeP}(B)$ is the nth draw from a Bernoulli process, with B defined as in (4.2), then

$$Z_n = \sum_{i=1}^{\infty} b_{ni}\delta_{\omega_i}\ , \qquad b_{ni} \sim \mathrm{Bernoulli}(\pi_i) \tag{4.3}$$

A set of N such draws, $\{Z_n\}_{n=1,N}$, may be used to define whether feature $\omega_i \in \Omega$ is utilized to represent the nth data sample, where $b_{ni} = 1$ if feature ω_i is employed, and $b_{ni} = 0$ otherwise. One may marginalize out the measure B analytically, yielding conditional probabilities for the $\{Z_n\}$ that correspond to the Indian buffet process (Thibaux and Jordan, 2007b; Griffiths and Ghahramani, 2005).

4.1.2 Covariate-dependent Lévy process

In the above beta-Bernoulli construction, the same measure $B \sim \mathrm{BP}(c, B_0)$ is employed for generation of all $\{Z_n\}$, implying that each of the N samples have the same probabilities $\{\pi_i\}$ for use of the respective features $\{\omega_i\}$. We now assume that with each of the N samples of interest there are an associated set of covariates, denoted respectively as $\{\boldsymbol{x}_n\}$, with each $\boldsymbol{x}_n \in \mathcal{X}$. We wish to impose that if samples n and n' have similar covariates $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$, that it is probable that they will employ a similar subset of the features $\{\omega_i\}$; if the covariates are distinct it is less probable that feature sharing will be manifested.

Generalizing (4.2), consider

$$\mathcal{B} = \sum_{i=1}^{\infty} \gamma_i \delta_{\omega_i}\ , \quad \omega_i \sim B_0 \tag{4.4}$$

where $\gamma_i = \{\gamma_i(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ is a stochastic process (random function) from $\mathcal{X} \to [0, 1]$ (drawn independently from the $\{\omega_i\}$). Hence, $\mathcal{B}$ is a dependent *collection* of Lévy processes with the measure specific to covariate $\boldsymbol{x} \in \mathcal{X}$ being $\mathcal{B}_{\boldsymbol{x}} = \sum_{i=1}^{\infty} \gamma_i(\boldsymbol{x})\delta_{\omega_i}$.

This constitutes a general specification, with several interesting special cases. For example, one might consider $\gamma_i(\boldsymbol{x}) = g\{\mu_i(\boldsymbol{x})\}$, where $g : \mathbb{R} \rightarrow [0, 1]$ is any monotone differentiable link function and $\mu_i(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathbb{R}$ may be modeled as a Gaussian process (Rasmussen and Williams, 2006), or related kernel-based construction. To choose $g\{\mu_i(\boldsymbol{x})\}$ one can potentially use models for the predictor-dependent breaks in probit, logistic or kernel stick-breaking processes (Rodriguez and Dunson, 2009; Ren et al., 2011b; Dunson and Park, 2008). In the remainder of this chapter we propose a special case for design of $\gamma_i(\boldsymbol{x})$, termed the *kernel* beta process (KBP).

4.1.3 Characteristic function of the kernel beta process

Recall from Hjort (Hjort, 1990) that $B \sim \mathrm{BP}(c(\omega), B_0)$ is a beta process on measure space $(\Omega, \mathcal{F})$ if its characteristic function satisfies

$$\mathbb{E}[e^{juB(\mathcal{A})}] = \exp\{\int_{[0,1]\times\mathcal{A}} (e^{ju\pi} - 1)\nu(d\pi, d\omega)\} \tag{4.5}$$

where here $j = \sqrt{-1}$, and $\mathcal{A}$ is any subset in $\mathcal{F}$. The beta process is a particular class of the Lévy process, with $\nu(d\pi, d\omega)$ defined as in (4.1).

For kernel $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*)$, let $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{x}^* \in \mathcal{X}$, and $\psi^* \in \Psi$; it is assumed that $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*) \in [0, 1]$ for all $\boldsymbol{x}$, $\boldsymbol{x}^*$ and ψ^*. As a specific example, for the radial basis function $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*) = \exp[-\psi^*\|\boldsymbol{x} - \boldsymbol{x}^*\|_2]$, where $\psi^* \in \mathbb{R}^+$. Let $\boldsymbol{x}^*$ represent random variables drawn from probability measure H, with support on $\mathcal{X}$, and ψ^* is also a random variable drawn from an appropriate probability measure Q with support over Ψ (*e.g.*, in the context of the radial basis function, ψ^* are drawn from a probability measure with support over $\mathbb{R}^+$). We now define a new Lévy measure

$$\nu_{\mathcal{X}} = H(d\boldsymbol{x}^*)Q(d\psi^*)\nu(d\pi, d\omega) \tag{4.6}$$

where $\nu(d\pi, d\omega)$ is the Lévy measure associated with the beta process, defined in

(4.1).

Theorem 1 Assume parameters $\{\boldsymbol{x}_i^*, \psi_i^*, \pi_i, \omega_i\}$ are drawn from measure $\nu_{\mathcal{X}}$ in (4.6), and that the following measure is constituted

$$\mathcal{B}_{\boldsymbol{x}} = \sum_{i=1}^{\infty} \pi_i K(\boldsymbol{x}, \boldsymbol{x}_i^*; \psi_i^*) \delta_{\omega_i} \tag{4.7}$$

which may be evaluated for *any* covariate $\boldsymbol{x} \in \mathcal{X}$. For any finite set of covariates $\mathcal{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{|\mathcal{S}|}\}$, we define the $|\mathcal{S}|$-dimensional random vector $\boldsymbol{K} = (K(\boldsymbol{x}_1, \boldsymbol{x}^*; \psi^*), \ldots, K(\boldsymbol{x}_{|\mathcal{S}|}, \boldsymbol{x}^*; \psi^*))^T$, with random variables $\boldsymbol{x}^*$ and ψ^* drawn from H and Q, respectively. For any set $\mathcal{A} \subset \mathcal{F}$, the $\mathcal{B}$ evaluated at covariates $\mathcal{S}$, on the set $\mathcal{A}$, yields an $|\mathcal{S}|$-dimensional random vector $\boldsymbol{\mathcal{B}}(\mathcal{A}) = (\mathcal{B}_{\boldsymbol{x}_1}(\mathcal{A}), \ldots, \mathcal{B}_{\boldsymbol{x}_{|\mathcal{S}|}}(\mathcal{A}))^T$, where $\mathcal{B}_{\boldsymbol{x}}(\mathcal{A}) = \sum_{i:\ \omega_i \in \mathcal{A}} \pi_i K(\boldsymbol{x}, \boldsymbol{x}_i^*; \psi_i^*)$. Expression (4.7) is a covariate-dependent Lévy process with Lévy measure (4.6), and characteristic function for an arbitrary set of covariates $\mathcal{S}$ satisfying

$$\mathbb{E}[e^{j<\boldsymbol{u}, \boldsymbol{\mathcal{B}}(\mathcal{A})>}] = \exp\{\int_{\mathcal{X} \times \Psi \times [0,1] \times \mathcal{A}} (e^{j<\boldsymbol{u}, \boldsymbol{K}\pi>} - 1) \nu_{\mathcal{X}}(d\boldsymbol{x}^*, d\psi^*, d\pi, d\omega)\} \tag{4.8}$$

□

A proof is provided in the Appendix. Additionally, for notational convenience, below a draw of (4.7), valid for all covariates in $\mathcal{X}$, is denoted $\mathcal{B} \sim \mathrm{KBP}(c, B_0, H, Q)$, with c and B_0 defining $\nu(d\pi, d\omega)$ in (4.1).

4.1.4 Relationship to the beta-Bernoulli process

If the covariate-dependent measure $\mathcal{B}_{\boldsymbol{x}}$ in (4.7) is employed to define covariate-dependent feature usage, then $Z_{\boldsymbol{x}} \sim \mathrm{BeP}(\mathcal{B}_{\boldsymbol{x}})$, generalizing (4.3). Hence, given $\{\boldsymbol{x}_i^*, \psi_i^*, \pi_i\}$, the feature-usage measure is $Z_{\boldsymbol{x}} = \sum_{i=1}^{\infty} b_{\boldsymbol{x}i} \delta_{\omega_i}$, with $b_{\boldsymbol{x}i} \sim \mathrm{Bernoulli}(\pi_i K(\boldsymbol{x}, \boldsymbol{x}_i^*; \psi_i^*))$. Note that it is equivalent in distribution to express $b_{\boldsymbol{x}i} = z_{\boldsymbol{x}i}^{(1)} z_{\boldsymbol{x}i}^{(2)}$, with $z_{\boldsymbol{x}i}^{(1)} \sim \text{Bernoulli}(K(\boldsymbol{x}, \boldsymbol{x}_i^*; \psi_i^*))$ and $z_{\boldsymbol{x}i}^{(2)} \sim \text{Bernoulli}(\pi_i)$. This model therefore yields the two-step generalization of the generative process of the beta-Bernoulli process discussed in the Introduction. The condition $z_{\boldsymbol{x}i}^{(1)} = 1$ only has a high probability when observed covariates $\boldsymbol{x}$ are near the (latent/inferred) covariates $\boldsymbol{x}_i^*$. It is deemed attractive that this intuitive generative process comes as a result of a rigorous Lévy process construction, the properties of which are summarized next.

4.1.5 Properties of $\mathcal{B}$

For all Borel subsets $\mathcal{A} \in \mathcal{F}$, if $\mathcal{B}$ is drawn from the KBP and for covariates $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, we have

$$
\begin{aligned}
\mathbb{E}[\mathcal{B}_{\boldsymbol{x}}(\mathcal{A})] &= B_0(\mathcal{A})\mathbb{E}(K_{\boldsymbol{x}}) \\
\text{Cov}(\mathcal{B}_{\boldsymbol{x}}(\mathcal{A}), \mathcal{B}_{\boldsymbol{x}'}(\mathcal{A})) &= \mathbb{E}(K_{\boldsymbol{x}} K_{\boldsymbol{x}'}) \int_{\mathcal{A}} \frac{B_0(d\omega)(1 - B_0(d\omega))}{c(\omega) + 1} - \text{Cov}(K_{\boldsymbol{x}}, K_{\boldsymbol{x}'}) \int_{\mathcal{A}} B_0^2(d\omega)
\end{aligned}
$$

where, $\mathbb{E}(K_{\boldsymbol{x}}) = \int_{\mathcal{X} \times \Psi} K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*) H(dx^*) Q(d\psi^*)$. If $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*) = 1$ for all $\boldsymbol{x} \in \mathcal{X}$, $\mathbb{E}(K_{\boldsymbol{x}}) = \mathbb{E}(K_{\boldsymbol{x}} K_{\boldsymbol{x}'}) = 1$, and $\text{Cov}(K_{\boldsymbol{x}}, K_{\boldsymbol{x}'}) = 0$, and the above results reduce to the those for the original BP (Thibaux and Jordan, 2007b).

Assume $c(\omega) = c$, where $c \in \mathbb{R}^+$ is a constant,
and let $\boldsymbol{K}_{\boldsymbol{x}} = (K(\boldsymbol{x}, \boldsymbol{x}_1^*; \psi_1^*), K(\boldsymbol{x}, \boldsymbol{x}_2^*; \psi_2^*), \dots)^T$ represent an infinite-dimensional vector, then for fixed kernel parameters $\{\boldsymbol{x}_i^*, \psi_i^*\}$,

$$
\text{Corr}(\mathcal{B}_{\boldsymbol{x}}(\mathcal{A}), \mathcal{B}_{\boldsymbol{x}'}(\mathcal{A})) = \frac{< \boldsymbol{K}_{\boldsymbol{x}}, \boldsymbol{K}_{\boldsymbol{x}'} >}{\|\boldsymbol{K}_{\boldsymbol{x}}\|_2 \cdot \|\boldsymbol{K}_{\boldsymbol{x}'}\|_2} \tag{4.9}
$$

where it is assumed $< \boldsymbol{K}_{\boldsymbol{x}}, \boldsymbol{K}_{\boldsymbol{x}'} >$, $\|\boldsymbol{K}_{\boldsymbol{x}}\|_2$, $\|\boldsymbol{K}_{\boldsymbol{x}'}\|_2$ are finite; the latter condition is always met when we (in practice) truncate the number of terms used in (4.7). The expression in (4.9) clearly imposes the desired property of high correlation in $\mathcal{B}_{\boldsymbol{x}}$ and

$\mathcal{B}_{\boldsymbol{x}'}$ when $\boldsymbol{x}$ and $\boldsymbol{x}'$ are proximate.

Proofs of the above properties are provided in the Appendix.

4.2 Applications

4.2.1 Model construction

We develop a covariate-dependent factor model, generalizing (Knowles and Ghahramani, 2007; Zhou et al., 2009), which did not consider covariates. Consider data $\boldsymbol{y}_n \in \mathbb{R}^M$ with associated covariates $\boldsymbol{x}_n \in \mathbb{R}^L$, with $n = 1, \ldots, N$. The factor loadings in the factor model here play the role of "dishes" in the buffet analogy, and we model the data as

$$\begin{gathered} \boldsymbol{y}_n = \mathbf{D}(\boldsymbol{w}_n \circ \boldsymbol{b}_n) + \boldsymbol{\epsilon}_n \\ Z_{\boldsymbol{x}_n} \sim \mathrm{BeP}(\mathcal{B}_{\boldsymbol{x}_n}), \qquad \mathcal{B} \sim \mathrm{KBP}(c, B_0, H, Q), \qquad B_0 \sim \mathrm{DP}(\alpha_0 G_0) \\ \boldsymbol{w}_n \sim \mathcal{N}(\mathbf{0}, \alpha_1^{-1}\mathbf{I}_T), \qquad \boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \alpha_2^{-1}\mathbf{I}_M) \end{gathered} \tag{4.10}$$

with gamma priors placed on α_0, α_1 and α_2, with $\circ$ representing the pointwise (Hadamard) vector product, and with $\mathbf{I}_M$ representing the $M \times M$ identity matrix. The Dirichlet process (Ferguson, 1973) base measure $G_0 = \mathcal{N}(0, \frac{1}{M}\mathbf{I}_M)$, and the KBP base measure B_0 is a *mixture* of atoms (factor loadings). For the applications considered it is important that the same atoms be reused at different points $\{\boldsymbol{x}_i^*\}$ in covariate space, to allow for repeated structure to be manifested as a function of space or time, within the image and music applications, respectively. The columns of $\mathbf{D}$ are defined respectively by $(\omega_1, \omega_2, \ldots)$ in $\mathcal{B}$, and the vector $\boldsymbol{b}_n = (b_{n1}, b_{n2}, \ldots)$ with $b_{nk} = Z_{\boldsymbol{x}_n}(\omega_k)$. Note that $\mathcal{B}$ is drawn *once* from the KBP, and when drawing the $Z_{\boldsymbol{x}_n}$ we evaluate $\mathcal{B}$ as defined by the respective covariate $\boldsymbol{x}_n$.

When implementing the KBP, we truncate the sum in (4.7) to T terms, and draw the $\pi_i \sim \mathrm{Beta}(1/T, 1)$, which corresponds to setting $c = 1$. We set T large, and the model infers the subset of $\{\pi_i\}_{i=1,T}$ that have significant amplitude, thereby estimating the number of factors needed for representation of the data. In practice we let H and Q be multinomial distributions over a discrete and finite set of, respectively, locations for $\{\boldsymbol{x}_i^*\}$ and kernel parameters for $\{\psi_i^*\}$, details of which are discussed in the specific examples.

In (4.10), the ith column of $\mathbf{D}$, denoted $\mathbf{D}_i$, is drawn from B_0, with B_0 drawn from a Dirichlet process (DP). There are multiple ways to perform such DP clustering, and here we apply the *Pólya urn scheme* (Ferguson, 1973). Assume $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{i-1}$ are a series of i.i.d. random draws from B_0, then the successive conditional distribution of $\mathbf{D}_i$ is of the following form:

$$\mathbf{D}_i | \mathbf{D}_1, \ldots, \mathbf{D}_{i-1}, \alpha_0, G_0 \sim \sum_{l=1}^{N_u} \frac{n_l^*}{i-1+\alpha_0} \delta_{\mathbf{D}_l^*} + \frac{\alpha_0}{i-1+\alpha_0} G_0, \tag{4.11}$$

where $\{\mathbf{D}_l^*\}_{l=1,N_u}$ are the unique dictionary elements shared by the first $i-1$ columns of $\mathbf{D}$, and $n_l^* = \sum_{j=1}^{i-1} \delta(\mathbf{D}_j = \mathbf{D}_l^*)$. For model inference, an indicator variable c_i is introduced for each $\mathbf{D}_i$, and $c_i = l$ with a probability proportional to n_l^*, with $l = 1, \ldots, N_u$, with c_i equal to $N_u + 1$ with a probability controlled by α_0. If $c_i = l$ for $l = 1, \ldots, N_u$, $\mathbf{D}_i$ takes the value $\mathbf{D}_l^*$; otherwise $\mathbf{D}_i$ is drawn from the prior $G_0 = \mathcal{N}(0, \frac{1}{M}\mathbf{I}_M)$, and a new dish/factor loading $\mathbf{D}_{N_u+1}^*$ is hence introduced.

4.2.2 Extensions

It is relatively straightforward to include additional model sophistication into (4.10), one example of which we will consider in the context of the image-processing example. Specifically, in many applications it is inappropriate to assume a Gaussian model for the noise or residual $\boldsymbol{\epsilon}_n$. In Section 4.3.3 we consider the following augmented noise model:

$$\boldsymbol{\epsilon}_n = \boldsymbol{\lambda}_n \circ \boldsymbol{m}_n + \hat{\boldsymbol{\epsilon}}_n \tag{4.12}$$

$$\boldsymbol{\lambda}_n \sim \mathcal{N}(\mathbf{0}, \alpha_\lambda^{-1}\mathbf{I}_M), \quad m_{np} \sim \text{Bernoulli}(\tilde{\pi}_n), \quad \tilde{\pi}_n \sim \text{Beta}(a_0, b_0), \quad \hat{\boldsymbol{\epsilon}}_n \sim \mathcal{N}(\mathbf{0}, \alpha_3^{-1}\mathbf{I}_M)$$

with gamma priors placed on α_λ and α_2, and with $p = 1, \ldots, M$. The term $\boldsymbol{\lambda}_n \circ \boldsymbol{m}_n$ accounts for "spiky" noise, with potentially large amplitude, and $\hat{\pi}_n$ represents the probability of spiky noise in data sample n. This type of noise model was considered in (Zhou et al., 2011), with which we compare.

4.2.3 Inference

The model inference is performed with a Gibbs sampler. Due to the limited space, only those variables having update equations distinct from those in the BP-FA of (Zhou et al., 2009) are included here. Assume T is the truncation level for the number of dictionary elements, $\{\mathbf{D}_i\}_{i=1,T}$; N_u is the number of unique dictionary elements values in the current Gibbs iteration, $\{\mathbf{D}_l^*\}_{l=1,N_u}$. For the applications considered in this chapter, $K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*)$ is defined based on the Euclidean distance: $K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*) = \exp[-\psi_i^* ||\boldsymbol{x}_n - \boldsymbol{x}_i^*||_2]$ for $i = 1, \ldots, T$; both ψ_i^* and $\boldsymbol{x}_i^*$ are updated from multinomial distributions (defining Q and H, respectively) over a set of discretized values with a uniform prior for each; more details on this are discussed in Section 4.3.

- **Update** $\{\mathbf{D}_l^*\}_{l=1,L}$: $\mathbf{D}_l^* \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$,

  $$\boldsymbol{\mu}_l = \boldsymbol{\Sigma}_l[\alpha_2 \sum_{n=1}^{N} \sum_{i:c_i=l} (b_{ni}w_{ni})\boldsymbol{y}_n^{-l}], \quad \boldsymbol{\Sigma}_l = [\alpha_2 \sum_{n=1}^{N} \sum_{i:c_i=l} (b_{ni}w_{ni})^2 + M]^{-1}\mathbf{I}_M,$$

  where $\boldsymbol{y}_n^{-l} = \boldsymbol{y}_n - \sum_{i:c_i \neq l} \mathbf{D}_i(b_{ni}w_{ni})$.

- **Update** $\{c_i\}_{i=1,T}$: $p(c_i) \sim \text{Mult}(\mathbf{p}_i)$,

$$p(c_i = l|-) \propto \begin{cases} \frac{n_l^{*-i}}{T-1+\alpha_0} \prod_{n=1}^N \exp\{-\frac{\alpha_2}{2}\|\boldsymbol{y}_n^{-i} - \mathbf{D}_l^*(b_{ni}w_{ni})\|_2^2\}, & \text{if } l \text{ is previously used,} \\ \frac{\alpha_0}{T-1+\alpha_0} \prod_{n=1}^N \exp\{-\frac{\alpha_2}{2}\|\boldsymbol{y}_n^{-i} - \mathbf{D}_{l^{new}}^*(b_{ni}w_{ni})\|_2^2\}, & \text{if } l = l^{new}, \end{cases}$$

where $n_l^{*-i} = \sum_{j:j\neq i} \delta(\mathbf{D}_j = \mathbf{D}_l^*)$, and $\boldsymbol{y}_n^{-i} = \boldsymbol{y}_n - \sum_{k:k\neq i} \mathbf{D}_k(b_{nk}w_{nk})$; $\mathbf{p}_i$ is realized by normalizing the above equation.

- **Update** $\{Z_{\boldsymbol{x}_n}\}_{n=1,N}$: for $Z_{\boldsymbol{x}_n}$, update each component $p(b_{ni}) \sim \text{Bernoulli}(v_{ni})$ for $i = 1, \ldots, K$,

$$\frac{p(b_{ni} = 1)}{p(b_{ni} = 0)} = \frac{\exp\{-\frac{\alpha_2}{2}\big[\mathbf{D}_i^T\mathbf{D}_i w_{ni}^2 - 2w_{ni}\mathbf{D}_i^T\boldsymbol{y}_n^{-i}\big]\}\pi_i K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*)}{1 - \pi_i K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*)}.$$

v_{ni} is calculated by normalizing $p(b_{ni})$ with the above constraint.

- **Update** $\{\pi_i\}_{i=1,T}$:

  Introduce two sets of auxiliary variables $\{z_{ni}^{(1)}\}_{i=1,T}$ and $\{z_{ni}^{(2)}\}_{i=1,T}$ for each data $\mathbf{y}_n$. Assume $z_{ni}^{(1)} \sim \text{Bernoulli}(\pi_i)$ and $z_{ni}^{(2)} \sim \text{Bernoulli}(K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*))$. For each specific n,

  – If $b_{ni} = 1$, $z_{ni}^{(1)} = 1$ and $z_{ni}^{(2)} = 1$;

  – If $b_{ni} = 0$, $\begin{cases} p(z_{ni}^{(1)} = 0, z_{ni}^{(2)} = 0|b_{ni} = 0) = \frac{(1-\pi_i)\big(1-K(\boldsymbol{x}_n,\boldsymbol{x}_i^*;\psi_i^*)\big)}{1-\pi_i K(\boldsymbol{x}_n,\boldsymbol{x}_i^*;\psi_i^*)} \\ p(z_{ni}^{(1)} = 0, z_{ni}^{(2)} = 1|b_{ni} = 0) = \frac{(1-\pi_i)K(\boldsymbol{x}_n,\boldsymbol{x}_i^*;\psi_i^*)}{1-\pi_i K(\boldsymbol{x}_n,\boldsymbol{x}_i^*;\psi_i^*)} \\ p(z_{ni}^{(1)} = 1, z_{ni}^{(2)} = 0|b_{ni} = 0) = \frac{\pi_i\big(1-K(\boldsymbol{x}_n,\boldsymbol{x}_i^*;\psi_i^*)\big)}{1-\pi_i K(\boldsymbol{x}_n,\boldsymbol{x}_i^*;\psi_i^*)} \end{cases}$

  From the above equations, we derive the conditional distribution for π_i,

$$\pi_i \sim \text{Beta}\big(\frac{1}{T} + \sum_n z_{ni}^{(1)}, 1 + \sum_n (1 - z_{ni}^{(1)})\big).$$

4.3 Experiments

4.3.1 Hyperparameter settings

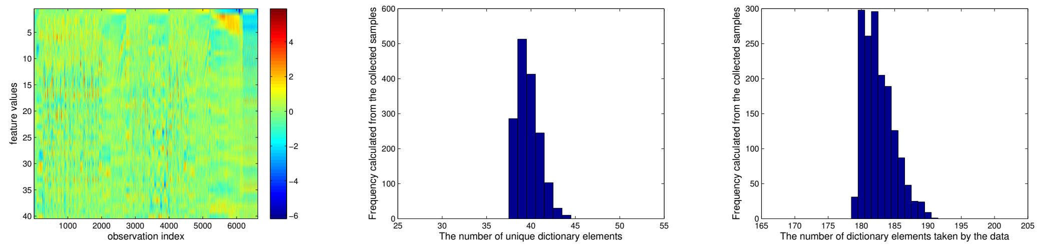
For both α_1 and α_2 the corresponding prior was set to Gamma$(10^{-6}, 10^{-6})$; the concentration parameter α_0 was given a prior Gamma$(1, 0.1)$. For both experiments below, the number of dictionary elements T was truncated to 256, the number of unique dictionary element values was initialized to 100, and $\{\pi_i\}_{i=1,T}$ were initialized to 0.5. All $\{\psi_i^*\}_{i=1,T}$ were initialized to 10^{-5} and updated from a set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ with a uniform prior Q. The remaining variables were initialized randomly. No parameter tuning or optimization has been performed.

4.3.2 Music analysis

We consider the same music piece as described in (Ren et al., 2010): "A Day in the Life" from the Beatles' album Sgt. Pepper's Lonely Hearts Club Band. The acoustic signal was sampled at 22.05 KHz and divided into 50 ms contiguous frames; 40-dimensional Mel frequency cepstral coefficients (MFCCs) were extracted from each frame, shown in Figure 4.1(a).

A typical goal of music analysis is to infer interrelationships within the music piece, as a function of time (Ren et al., 2010). For the audio data, each MFCC vector $\boldsymbol{y}_n$ has an associated time index, the latter used as the covariate $\boldsymbol{x}_n$. The finite set of temporal sample points (covariates) were employed to define a library for the $\{\boldsymbol{x}_i^*\}$, and H is a uniform distribution over this set. After 2000 burn-in iterations, we collected samples every five iterations. Figure 4.1(b) shows the frequency for the number of *unique* dictionary elements used by the data, based on the 1600 collected samples; and Figure 4.1(c) shows the frequency for the number of total dictionary elements used.

With the model defined in (4.10), the sparse vector $\boldsymbol{b}_n \circ \mathbf{w}_n$ indicates the impor-

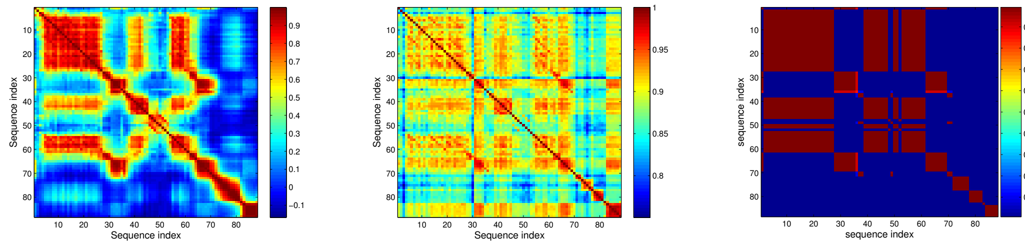


(a) (b) (c)

FIGURE 4.1: (a) MFCCs features used in music analysis, where the horizontal axis corresponds to time, for "A Day in the Life". Based on the Gibbs collection samples: (b) frequency on number of *unique* dictionary elements, and (c) *total* number of dictionary elements.

tance of each dictionary element from $\{\mathbf{D}_i\}_{i=1,T}$ to data $\boldsymbol{y}_n$. Each of these N vectors $\{\boldsymbol{b}_n \circ \mathbf{w}_n\}_{n=1,N}$ was normalized within each Gibbs sample, and used to compute a correlation matrix associated with the N time points in the music. Finally, this matrix was averaged across the collection samples, to yield a correlation matrix relating one part of the music to all others. For a fair comparison between our methods and the model proposed in (Ren et al., 2010) (which used an HMM, and computed correlations over *windows* of time), we divided the whole piece into multiple consecutive short-time windows. Each temporal window includes 75 consecutive feature vectors, and we compute the average correlation coefficients between the features within each pair of windows. There were 88 temporal windows in total (each temporal window is denoted as a sequence in Figure 4.2), and the dimension of the correlation matrix is accordingly 88×88. The computed correlation matrix for the proposed KBP model is presented in Figure 4.2(a).

We compared KBP performance with results based on BP-FA (Zhou et al., 2009) in which covariates are not employed, and with results from the dynamic clustering model in (Ren et al., 2010), in which a dynamic HMM is employed (in (Ren et al., 2010) a *dynamic* HDP, or dHDP, was used in concert with an HMM). The BP-FA results correspond to replacing the KBP with a BP. The correlation matrix computed



(a) (b) (c)

FIGURE 4.2: Inference of relationships in music as a function of time, as computed via a correlation of the dictionary-usage weights, for (a) and (b), and based upon state usage in an HMM, for (c). Results are shown for "A Day in the Life." The results in (c) are from (Ren et al., 2010), as a courtesy from the authors of that paper. (a) KBP-FA, (b) BP-FA, (c) dHDP-HMM .

from the BP-FA and the dHDP-HMM (Ren et al., 2010) are shown in Figures 4.2(b) and (c), respectively. The dHDP-HMM results yield a reasonably good segmentation of the music, but it is unable to infer subtle differences in the music over time (for example, all voices in the music are clustered together, even if they are different). Since the BP-FA does not capture as much localized information in the music (the probability of dictionary usage is the same for all temporal positions), it does not manifest as good a music segmentation as the dHDP-HMM. By contrast, the KBP-FA model yields a good music segmentation, while also capturing subtle differences in the music over time (*e.g.*, in voices). Note that the use of the DP to allow repeated use of dictionary elements as a function of time (covariates) is important here, due to the repetition of structure in the piece. One may listen to the music and observe the segmentation at `http://www.youtube.com/watch?v=35YhHEbIlEI`.

4.3.3 Image interpolation and denoising

We consider image interpolation and denoising as two additional potential applications. In both of these examples each image is divided into N 8×8 overlapping patches, and each patch is stacked into a vector of length $M = 64$, constituting observation $\boldsymbol{y}_n \in \mathbb{R}^M$. The covariate $\boldsymbol{x}_n$ represents the patch coordinates in the 2-D space. The probability measure H corresponds to a uniform distribution over the centers of all 8×8 patches. The images were recovered based on the average of the collection samples, and each pixel was averaged across all overlapping patches in which it resided. For the image-processing examples, 5000 Gibbs samples were run, with the first 2000 discarded as burn-in.

For image interpolation, we only observe a fraction of the image pixels, sampled uniformly at random. The model infers the underlying dictionary $\mathbf{D}$ in the presence of this missing data, as well as the weights on the dictionary elements required for representing the observed components of $\{\boldsymbol{y}_n\}$; using the inferred dictionary and associated weights, one may readily impute the missing pixel values. In Table 4.1 we present average PSNR values on the recovered pixel values, as a function of the fraction of pixels that are observed (20% in Table 4.1 means that 80% of the pixels are missing uniformly at random). Comparisons are made between a model based on BP and one based on the proposed KBP; the latter generally performs better, particularly when a large fraction of the pixels are missing. The proposed algorithm yields results that are comparable to those in (Zhou et al., 2011), which also employed covariates within the BP construc tion. However, the proposed KBP construction has the significant computational advantages of only requiring kernels centered at the locations of the dictionary-dependent covariates $\{\boldsymbol{x}_i^*\}$, while the model in (Zhou et al., 2011) has a kernel for each of the image patches, and therefore it scales unfavorably for large images.

In the image-denoising example in Figure 4.3 the images were corrupted with both white Gaussian noise (WGN) and sparse spiky noise, as considered in (Zhou et al., 2011). The sparse spiky noise exists in particular pixels, selected uniformly at random, with amplitude distributed uniformly between -255 and 255. For the pepper image, 15% of the pixels were corrupted by spiky noise, and the standard

Table 4.1: Comparison of BP and KBP for interpolating images with pixels missing uniformly at random, using standard image-processing images. The top and bottom rows of each cell show results of BP and KBP, respectively. Results are shown when 20%, 30% and 50% of the pixels are observed, selected uniformly at random.

RATIO	C.MAN	HOUSE	PEPPERS	LENA	BARBARA	BOATS	F.PRINT	MAN	COUPLE	HILL
20%	23.75	29.75	25.56	30.97	26.84	27.84	26.49	28.29	27.76	29.38
	24.02	**30.89**	**26.29**	**31.38**	**28.93**	**28.11**	**26.89**	**28.37**	**28.03**	**29.67**
30%	25.59	33.09	28.64	33.30	30.13	30.20	29.23	29.89	29.97	31.19
	25.75	**34.02**	**29.29**	**33.33**	**31.46**	**30.24**	**29.37**	**30.12**	**30.33**	**31.25**
50%	28.66	38.26	32.53	**36.79**	35.95	33.05	**33.50**	**33.19**	**33.61**	**34.19**
	28.78	**38.35**	**32.69**	35.89	**36.03**	**33.18**	32.18	32.35	32.35	32.60

deviation of the WGN was 15; for the house image, 10% of the pixels were corrupted by spiky noise and the standard deviation of WGN was 10. We compared with different methods on both two images: the augmented KBP-FA model (KBP-FA+) in Section 4.2.2, the BP-FA model augmented with a term for spiky noise (BP-FA+) and the original BP-FA model. The model proposed with KBP showed the best denoising result for both visual and quantitative evaluations. Again, these results are comparable to those in (Zhou et al., 2011), with the significant computational advant age discussed above. Note that here the imposition of covariates and the KBP yields marked improvements in this application, relative to BP-FA alone.

4.4 Summary

A new Lévy process, the kernel beta process, has been developed for the problem of nonparametric Bayesian feature learning, with example results presented for music analysis, image denoising, and image interpolation. In addition to presenting theoretical properties of the model, state-of-the-art results are realized on these learning tasks. The inference is performed via a Gibbs sampler, with analytic update equations. Concerning computational costs, for the music-analysis problem, for example, the BP model required around 1 second per Gibbs iteration, with KBP requiring about 3 seconds, with results run on a PC with 2.4GHz CPU, in non-optimized

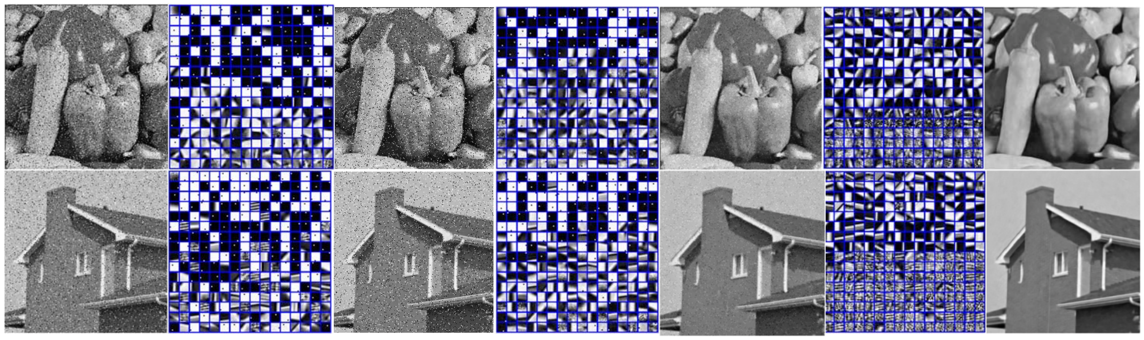


FIGURE 4.3: Denoising Result: the first column shows the noisy images (PSNR is 15.56 dB for Peppers and 17.54 dB for House); the second and third column shows the results inferred from the BP-FA model (PSNR is 16.31 dB for Peppers and 17.95 dB for House), with the dictionary elements shown in column two and the reconstruction in column three; the fourth and fifth columns show results from BP-FA+ (PSNR is 23.06 dB for Peppers and 26.71 dB for House); the sixth and seventh column shows the results of the KBP-FA+ (PSNR is 27.37 dB for Peppers and 34.89 dB for House). In each case the dictionaries are ordered based on their frequency of usage, starting from top-left.

Matlab™.

5

Scalable Bayesian Low-Rank Tensor Representation

In the previous chapters, our study about the stochastic processes in the applications of nonparametric Bayesian methods is focused on the scope of Lévy processes - the Lévy measure decomposition in Chapter 3 and kernel beta process in Chapter 4. While in this chapter, we apply a non-Lévy process, the multiplicative gamma process (Bhattacharya and Dunson, 2011) in the low-rank representation of tensors for the multiway data inference task. To be specific, the multiplicative gamma process prior is applied along the super-diagonal in the CANDECOMP/PARAFAC (CP) decomposition (Kolda and Bader, 2009) of tensors to infer the tensor ranks in a nonparametric way.

5.1 Low-Rank Tensor Decomposition

In this section, we present our framework for low-rank tensor decomposition based on the CP decomposition (Kolda and Bader, 2009). We infer the rank by placing a shrinkage prior, the multiplicative gamma process (MGP) (Bhattacharya and Dunson, 2011), over the superdiagonal elements of the *core tensor* (Λ in Figure 5.1) in the CP decomposition. The MGP prior adaptively learns the appropriate number of component tensors, and leads to an efficient low-rank approximation of the tensor.

5.1.1 CP Decomposition of Tensor

The CANDECOMP/PARAFAC (CP) decomposition decomposes a tensor into a sum of rank-1 *component tensors* (Kolda and Bader, 2009). A K-way (or K-mode) tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$, with the integer n_k being the dimension of $\mathcal{X}$ along the k^{th} way, can be represented in its CP decomposition form:

$$\mathcal{X} = \sum_{r=1}^{R} \lambda_r \cdot \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \cdots \circ \boldsymbol{u}_r^{(K)} \tag{5.1}$$

where the vector $\boldsymbol{u}_r^{(k)} \in \mathbb{R}^{n_k}$ and '$\circ$' denotes the vector outer product. Here R is referred to as the *rank* of the tensor $\mathcal{X}$. With the CP decomposition as given in (5.1), the tensor element $x_{\boldsymbol{i}}$, with $\boldsymbol{i} = [i_1, i_2, \cdots, i_K]$ its K-dimensional index vector, can be concisely represented by:

$$x_{\boldsymbol{i}} = \sum_{r=1}^{R} \lambda_r \prod_{k=1}^{K} u_{i_k r}^{(k)} \tag{5.2}$$

Denote by $U^{(k)} = [\boldsymbol{u}_1^{(k)}, \boldsymbol{u}_2^{(k)}, \cdots, \boldsymbol{u}_R^{(k)}]$, $k = 1, 2, \cdots, K$, the $n_k \times R$ *factor matrix* of the k-th mode of the tensor. When a vector form of the tensor $\mathcal{X}$ is desired, the above CP decomposition can be written as:

$$vec(\mathcal{X}) = U^{(1)} \odot U^{(2)} \odot \cdots \odot U^{(K)} \cdot \boldsymbol{\lambda} \tag{5.3}$$

where $\odot$ denotes the Khatri-Rao product and $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_R]$ denotes the vector along the superdiagonal of the core tensor.

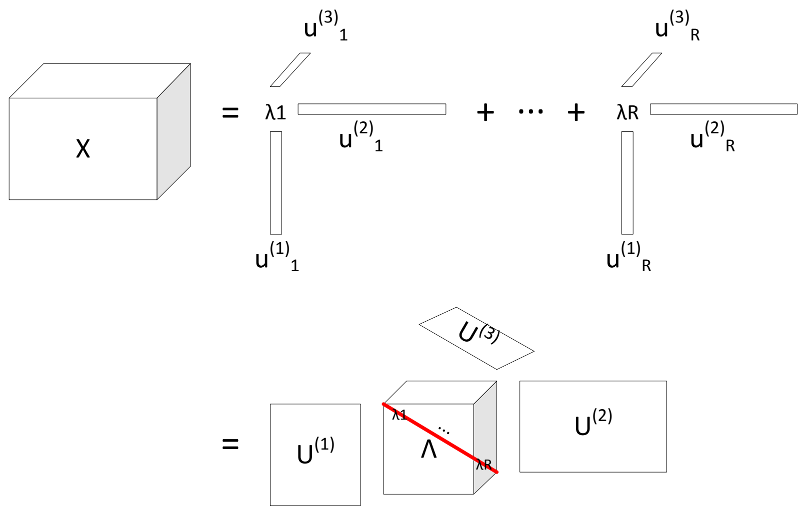


FIGURE 5.1: The CP decomposition of tensors (a three-mode tensor shown for illustration).

5.1.2 Rank Specification

The CP decomposition yields a concise representation of tensors. However, the trade-off of the conciseness is that in the CP decomposition the rank of the tensor being decomposed needs to be pre-specified. However, rank estimation for tensors is in general an NP hard problem (Hastad, 1990). To avoid the burdensome rank estimation task, a reasonable solution is to express the original tensor with a "good-enough" low-rank approximation; for example, in the sense of the Frobenius norm. But unfortunately, unlike the 2-way (matrix) cases, where the low-rank approximation is completely solved with the Eckart-Young theorem (Eckart and Young, 1936), for tensors the low-rank approximation can often be an ill-posed problem as discussed in (de Silva and Lim, 2008).

Such theoretical dilemma inspires alternative solutions for low-rank approximation of tensors. Instead of relying on ad-hoc or cumbersome model selection methods such as AIC, BIC, or the marginal likelihood, we turn to the nonparametric Bayesian modeling paradigm to adaptively infer the rank of the tensor being decomposed (or a close approximation of the rank necessary to obtain a sufficiently good low-rank approximation for a given dataset). In particular, we propose a nonparametric Bayesian low-rank CP decomposition for tensors based on the theoretically well motivated multiplicative gamma process (Bhattacharya and Dunson, 2011) prior (MGP) construction to infer the rank, as opposed to priors such as the Indian Buffet Process (Griffiths and Ghahramani, 2011) for which inference can be complicated/slow. As we show subsequently, the shrinkage property of the MGP leads to fully conjugate models in both continuous and binary data cases and allows us to derive simple, closed-form Gibbs sampling updates for all model parameters. For the binary case in particular, the conjugacy is achieved via the Pólya-Gamma sampling strategy (Polson et al., 2012) which elicits a closed-form Gibbs sampler.

5.1.3 CP Decomposition with MGP

Our low-rank tensor decomposition model construction is based on the multiplicative gamma process (MGP), originally proposed in the context of factor analysis of matrix data (Bhattacharya and Dunson, 2011). In (Bhattacharya and Dunson, 2011), this prior was employed on the columns of the factor loading matrix, such that the columns increasingly shrink to zero as the column index increases. We generalize this construction for the multi-way tensor case. Crucially, different from the construction used in (Bhattacharya and Dunson, 2011), for the low-rank decomposition of tensors we put the MGP prior on the superdiagonal elements $\boldsymbol{\lambda}$ of the core tensor Λ. This greatly reduces the number of parameters to be estimated in the tensor case. We denote this CP decomposition driven by the MGP as MGP-CP. The MGP prior is represented by:

$$
\begin{aligned}
&\lambda_r \sim \mathcal{N}(0, \tau_r^{-1}), \quad 1 \leqslant r \leqslant R \\
&\tau_r = \prod_{l=1}^{r} \delta_l, \quad \delta_l \sim Ga(a_c, 1) \quad a_c > 1
\end{aligned} \tag{5.4}
$$

The multiplicative gamma process prior described in (5.4) on the precision of the Gaussian distribution for λ_r will shrink the λ_r towards zero as r increases. The appropriate rank R under our MGP based tensor decomposition model can be inferred two ways: (i) using a reasonably large truncation level, or (ii) using an adaptation strategy (discussed in Section 5.2.2) which allows growing or shrinking ranks as inference progresses. We refer to the truncation-based version as MGP-CP^t and the adaptation-based version as MGP-CP^a.

We assume that for each mode of the tensor, the R columns $\boldsymbol{u}_r^{(k)}$ of the factor matrix $U^{(k)}$, are drawn from a Gaussian distribution:

$$
\boldsymbol{u}_r^{(k)} \sim N(\boldsymbol{\mu}^{(k)}, \Sigma^{(k)}),\ 1 < r \leqslant R,\ 1 < k \leqslant K \tag{5.5}
$$

where $\boldsymbol{\mu}^{(k)}$ and $\Sigma^{(k)}$ are the mean vector and covariance matrix of the Gaussian distribution of the k^{th} tensor mode. Then the covariance between any two elements in the tensor $\mathcal{X}$, $x_{\boldsymbol{i}}$ and $x_{\boldsymbol{j}}$ conditioned on the factor matrices is:

$$
\mathrm{Cov}(x_{\boldsymbol{i}}, x_{\boldsymbol{j}} | \{U^{(k)}\}_{k=1}^K) = \sum_{r=1}^{R} \tau_r^{-1} \prod_{k=1}^{K} u_{i_k r}^{(k)} u_{j_k r}^{(k)} \tag{5.6}
$$

In (5.6) we observe that the covariance is structured as the sum of the covariances associated with each rank-one component tensor, indicating that each component tensor stands for an independent significant factor constituting the data.

5.2 Model and Inference

5.2.1 Model Description

The goal of inference in our model is to infer the parameters of the CP decomposition, Λ, $U^{(1)}, U^{(2)}, \cdots, U^{(K)}$, based on potentially a very limited (sparse) set of observations $\mathcal{Y} = \{y_{\boldsymbol{i}}\}_{\boldsymbol{i} \in I}$, where I is the index set of all the observations, and $N = |\mathcal{Y}|$ the number of these observations. Following the MGP-CP model given in (5.4) and (5.5), the prior, $p(\Lambda, \{U^{(k)}\}_{k=1}^{K})$, is given below

$$\prod_{r=1}^{R} \mathcal{N}(\lambda_r | 0, \tau_r^{-1}) \mathrm{Ga}(\delta_r | a_r, 1) \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{u}_r^{(k)} | \boldsymbol{\mu}_r^{(k)}, \Sigma_r^{(k)}) \tag{5.7}$$

We further assume that the covariance matrices $\Sigma_r^{(k)}$'s are diagonal, which amounts to assuming that the entities in each tensor mode are *a priori* independent of each other.

Two types of likelihood models are considered based on different types of real-world data: continuous and binary data. The observations $\mathcal{Y}$ are assumed to be i.i.d. For the continuous observations with Gaussian noise, where τ_ϵ is the precision, the model likelihood is given by

$$p(\mathcal{Y}|\mathcal{X}) = \prod_{\boldsymbol{i}} \mathcal{N}(y_{\boldsymbol{i}} | x_{\boldsymbol{i}}, \tau_\epsilon^{-1}) \tag{5.8}$$

For binary-valued data (*e.g.*, relational data) the logistic link function is applied:

$$p(\mathcal{Y}|\mathcal{X}) = \prod_{\boldsymbol{i}} (\frac{1}{1 + \mathrm{e}^{-x_{\boldsymbol{i}}}})^{y_{\boldsymbol{i}}} (\frac{\mathrm{e}^{-x_{\boldsymbol{i}}}}{1 + \mathrm{e}^{-x_{\boldsymbol{i}}}})^{1 - y_{\boldsymbol{i}}} \tag{5.9}$$

5.2.2 Inference via Gibbs Sampling

For continuous data, our model construction with prior in (5.7) and likelihood model in (5.8) is locally conjugate and a Gibbs sampler can easily be derived for all the model parameters. For the binary case, the logistic likelihood in (5.9) is not conjugate to the prior in (5.7). To achieve conjugacy in the binary case, we use the Pólya-Gamma sampling strategy (Polson et al., 2012), which allows us to derive a fully analytic Gibbs sampler in the binary case as well.

The Gibbs sampling update equations for the various model parameters $\{\delta_r\}_{r=1}^R$, $\{\lambda_r\}_{r=1}^R$, and $\{U^{(k)}\}_{k=1}^K$, given continuous or binary observations $\mathcal{Y}$, are as follows:

(i) For the update of the MGP, for $\delta_r, 1 \leqslant r \leqslant R$:

$$\delta_r \sim \mathrm{Ga}(a_c + \frac{1}{2}(R - r + 1), 1 + \frac{1}{2}\sum_{h=r}^{R} \lambda_h^2 \prod_{l=1, l\neq r}^{h} \delta_l) \tag{5.10}$$

(ii) When the observations $\mathcal{Y}$ are real and the likelihood is given as in (5.8), for the update of $\lambda_r, 1 \leqslant r \leqslant R$:

$$x_{\boldsymbol{i}} = (\prod_{k=1}^{K} u_{i_k r}^{(k)})\lambda_r + (\sum_{r' \neq r} \lambda_{r'} \prod_{k=1}^{K} u_{i_k r'}^{(k)}) = a_{\boldsymbol{i}}^r \lambda_r + b_{\boldsymbol{i}}^r \tag{5.11}$$

$$\begin{aligned} &\lambda_r \sim \mathcal{N}(\hat{\mu}_r, \hat{\tau}_r^{-1}), \quad \hat{\tau}_r = \tau_r + \tau_\epsilon \sum_{\boldsymbol{i}} a_{\boldsymbol{i}}^{r\,2} \\ &\hat{\mu}_r = \hat{\tau}_r^{-1} \tau_\epsilon \sum_{\boldsymbol{i}} a_{\boldsymbol{i}}^r (y_{\boldsymbol{i}} - b_{\boldsymbol{i}}^r) \end{aligned} \tag{5.12}$$

For the update of $\boldsymbol{u}_r^{(k)}, 1 \leqslant r \leqslant R, 1 \leqslant k \leqslant K$, denote:

$$\begin{aligned} x_{\boldsymbol{i}} &= (\lambda_r \prod_{k' \neq k} u_{\boldsymbol{i}_{k'} r}^{(k)}) u_{\boldsymbol{i}_k r}^{(k)} + (\sum_{r' \neq r} \lambda_{r'} \prod_{k=1}^{K} u_{\boldsymbol{i}_k r'}^{(k)}) \\ &= c_{\boldsymbol{i}_k r}^{(k)} u_{\boldsymbol{i}_k r}^{(k)} + d_{\boldsymbol{i}_k r}^{(k)} \end{aligned} \tag{5.13}$$

Then $\boldsymbol{u}_r^{(k)} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_r^{(k)}, \hat{\Sigma}_r^{(k)})$ with:

$$\begin{aligned} \hat{\Sigma}_r^{(k)} &= ({\Sigma^{(k)}}^{-1} + T_r^{(k)})^{-1} \\ T_r^{(k)} &= \operatorname{diag}(\tau_{1r}^{(k)}, \tau_{2r}^{(k)}, \cdots, \tau_{n_k r}^{(k)}) \\ \tau_{nr}^{(k)} &= \tau_\epsilon \sum_{\boldsymbol{i}, i_k = n} {c_{i_k r}^{(k)}}^2, \quad 1 \leqslant n \leqslant n_k \end{aligned} \tag{5.14}$$

$$\begin{aligned} \hat{\boldsymbol{\mu}}_r^{(k)} &= \hat{\Sigma}_r^{(k)} ({\Sigma^{(k)}}^{-1} \boldsymbol{\mu}^{(k)} + T_r^{(k)} \boldsymbol{\alpha}_r^{(k)}) \\ \boldsymbol{\alpha}_r^{(k)} &= [\alpha_{1r}^{(k)}, \alpha_{2r}^{(k)}, \cdots, \alpha_{n_k r}^{(k)}]^\top, \text{for } 1 \leqslant n \leqslant n_k : \\ \alpha_{nr}^{(k)} &= (\tau_{nr}^{(k)})^{-1} \tau_\epsilon \sum_{\boldsymbol{i}, i_k = n} c_{i_k r}^{(k)} (y_{\boldsymbol{i}} - d_{i_k r}^{(k)}) \end{aligned} \tag{5.15}$$

Additionally we put a gamma prior on the noise precision $\tau_\epsilon \sim Ga(a_0, b_0)$, with the posterior $\hat{\tau}_\epsilon \sim Ga(a_0 + \frac{1}{2}N, b_0 + \frac{1}{2}\sum_{\boldsymbol{i}}(x_{\boldsymbol{i}} - \hat{x}_{\boldsymbol{i}})^2)$, with $\hat{x}_{\boldsymbol{i}}$ is the estimation of $x_{\boldsymbol{i}}$ reconstructed by following (5.2).

(iii) When the observations $\mathcal{Y}$ are binary and the likelihood is given as in (5.9), the model is a latent Gaussian model (LGM) with Logit likelihood. We apply the recent result of (Polson et al., 2012) which elicits a conjugate Gibbs sampler. For the update of $\lambda_r, 1 \leqslant r \leqslant R$, with (5.11) we have, $\lambda_r = \frac{1}{a_{\boldsymbol{i}}^r} x_{\boldsymbol{i}} - \frac{b_{\boldsymbol{i}}^r}{a_{\boldsymbol{i}}^r}$. Then the augment random variable $\phi_{\boldsymbol{i}}$ are drawn independently from the Pólya-Gamma distribution:

$$\phi_{\boldsymbol{i}} \sim \mathrm{PG}(1, x_{\boldsymbol{i}}) \tag{5.16}$$

where $\mathrm{PG}(\cdot,\cdot)$ represents the Pólya-Gamma distribution. Then $\lambda_r, 1 \leqslant r \leqslant R$ is drawn from Gaussian:

$$\begin{aligned} \lambda_r &\sim \mathcal{N}(\hat{\mu}_r, \hat{\tau}_r^{-1}), \quad \hat{\tau}_r = \tau_r + \sum_{\boldsymbol{i}} a_{\boldsymbol{i}}^{r\,2} \phi_{\boldsymbol{i}} \\ \hat{\mu}_r &= \hat{\tau}_r^{-1} \sum_{\boldsymbol{i}} a_{\boldsymbol{i}}^r (y_{\boldsymbol{i}} - 0.5 - \phi_{\boldsymbol{i}} b_{\boldsymbol{i}}^r) \end{aligned} \tag{5.17}$$

For the update of $\boldsymbol{u}_r^{(k)}, 1 \leqslant r \leqslant R, 1 \leqslant k \leqslant K$, with (5.13) we have: $u_{\boldsymbol{i}_k r}^{(k)} = \frac{1}{c_{\boldsymbol{i}_k r}^{(k)}} x_{\boldsymbol{i}} - \frac{d_{\boldsymbol{i}_k r}^{(k)}}{c_{\boldsymbol{i}_k r}^{(k)}}$. Then $\boldsymbol{u}_r^{(k)} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_r^{(k)}, \hat{\Sigma}_r^{(k)})$ with the $\hat{\boldsymbol{\mu}}_r^{(k)}, \hat{\Sigma}_r^{(k)}$ given as same as in (5.14) and (5.15), but $\tau_{nr}^{(k)}$, $\alpha_{nr}^{(k)}$ changed. For $1 \leqslant n \leqslant n_k$:

$$\begin{aligned} \tau_{nr}^{(k)} &= \sum_{\boldsymbol{i}, i_k = n} {c_{i_k r}^{(k)}}^2 \phi_{\boldsymbol{i}}, \quad 1 \leqslant n \leqslant n_k \\ \alpha_{nr}^{(k)} &= (\tau_{nr}^{(k)})^{-1} \sum_{\boldsymbol{i}, i_k = n} c_{i_k r}^{(k)} (y_{\boldsymbol{i}} - 0.5 - \phi_{\boldsymbol{i}} d_{i_k r}^{(k)}) \end{aligned} \tag{5.18}$$

Adaptation Strategy: In our truncation based variant, MGP-CP^t, we run the Gibbs sampler using a reasonably large truncation level R and, as inference progresses, only relevant components have significant contribution to the model, with the λ_r for the rest shrinking to values close to zero. In our adaptive variant, MGP-CP^a, whenever λ_r becomes smaller than a predefined threshold th (say 0.001), the component tensors with $|\lambda_r| < th$ are removed from the model; otherwise if all the $|\lambda_r| > th$, a new component tensor is added. Such adaptation occurs with probability $p(t) = \exp(\beta_0 + \beta_1 t)$ at the t^{th} iteration, with β_0, β_1 chosen so that adaptation occurs around every 10 iterations at the beginning of the chain but decreases in frequency exponentially fast (Bhattacharya and Dunson, 2011). The simple strategy of thresholding based on the absolute values of λ_r worked well in all of our experiments. Other criteria can also be used to decide whether to discard a rank-1 component or to add a new component. For example, in the continuous -data case, one possible strategy would be to monitor the explained variances by each rank-1 component on some held-out data and if the contribution of certain rank-1 components to the total explained variance drops below a very small value (say <1% of the explained variance), we drop them (otherwise we add a new component based on the adaptation probability $p(t)$). In the binary data case, we can likewise monitor the contributions by each rank-1 component to the predictive probabilities of all the observations and drop components which are *non-informative*, *e.g.*, if the empirical distribution estimated using the predictive probabilities of all the observations is close to a uniform distribution *and* if the mean of the empirical distribution is close to 0.5 (otherwise we add a new component based on $p(t)$).

5.2.3 Computational Complexity

The per-iteration computational cost of our inference algorithm is *linear* in the number of observation N. For sparse tensors N is considerably smaller than the tensor size $L = \prod_{k=1}^{K} n_k$. The individual contributions to the overall time complexity are as follows: (*i*) sampling each δ_r, $r = \{1, \ldots, R\}$ takes $O(R^2)$ time leading to a time-complexity $O(R^3)$; (*ii*) sampling each λ_r, $r = \{1, \ldots, R\}$ takes $O(NK)$ time leading to a time-complexity $O(NRK)$; (*iii*) sampling each $\boldsymbol{u}_r^{(k)}$, $r = \{1, \ldots, R\}, k = \{1, \ldots, K\}$ takes $O(NRK)$ time leading to a time-complexity $O(NK^2R^2)$. Note that no explicit matrix inversions are involved in our inference procedure since the covariance of the prior on $\boldsymbol{u}_r^{(k)}$ is assumed to be diagonal and therefore the computations in (5.14) can be performed in $O(n_k)$ time. The overall time-complexity is dominated by the third term $O(NR^2K^2)$ which is linear in the number of observations N. This is especially encouraging for a sampling based inference method.

The linear scalability of our method in N is appealing since real-world tensor datasets tend to be extremely sparse ($N \ll L$). For example, in most social network datasets, there are less than 0.1% observed interactions. Our experiments corroborate the linear scalability behavior (Section 5.4.5, Figure 5.6) on a sparsely observed tensor of size $1000 \times 1000 \times 1000$ for which L is 1 billion.

5.3 Related Work

With the advent of social networks and multirelational/multiway data observed in many application domains, tensor decomposition methods have gained much attention recently. Although a number of tensor decomposition methods have been proposed, many state-of-the-art methods (Nickel et al., 2011; Bordes et al., 2012; Jenatton et al., 2012) are specialized for analyzing three-mode tensor data and do not generalize to higher-order tensors.

Tensor decomposition methods that can infer the rank are relatively few. Among the probabilistic approaches, one option is to use the Automatic Relevance Determination (ARD) method (Mørup and Hansen, 2009; Zhao et al., 2014). We use this method as a baseline in our experiments. Some non-probabilistic methods for tensor decomposition employ trace-norm regularization (Tomioka et al., 2010) to get an approximation of the tensor rank. In another recent work, nuclear-norm based rank-regularization (Bazerque et al., 2013) is used to infer the rank in a probabilistic tensor factorization model and inference is based on MAP estimation. All of these methods assume that the observations are real-valued unlike our method which can deal with both real and binary data.

A nonparametric Bayesian method similar in spirit to ours is presented in (Dunson and Xing, 2012), where a stick-breaking prior is put on the superdiagonal of Λ in the CP decomposition of a probability tensor (a special type of tensor whose entries sum to 1), to nonparametrically learn a low-rank representation through inference. The main consideration for applying the stick-breaking process prior in (Dunson and Xing, 2012) comes from its statistically decreasing property and the norm one requirement of the specific type of tensors (three-mode probability tensor). In another recent work, (Yoshii et al., 2013) proposed a positive semidefinite tensor factorization (PSDTF) which corresponds to the CP decomposition where the rank is learned with a truncation level put on the gamma random variables along the superdiagonal of Λ. However, this method is also limited to special types of 3-way tensors where each slice is a positive semidefinite matrix. A potential alternative is to put a gamma process along the superdiagonal is the construction of the gamma process discussed in (Wang and Carin, 2012), where the statistically decreasing samples facilitate the nonparametric learning of the required rank. However, the resulting model will not conjugate with this choice of the prior.

Tensor decomposition methods that explicitly model binary data are also relatively few. The recently proposed Infinite Tucker Decomposition method (Xu et al., 2013) uses a probit model for binary data. We compare with this method in our experiments. Another recent work on modeling binary tensor data is a logistic loss based extension of the non-probabilistic RESCAL model (Nickel and Tresp, 2013). This is, however, limited to three-mode tensor data.

5.4 Experiments

We perform experiments with our model on both synthetic and real-world tensor datasets, and compare it with several baselines. The datasets used in our experiments span a wide range of application domains, such as chemometrics, multirelational social networks, brain-signal analysis (EEG), and image analysis. We experiment with both variants of our model: truncated MGP (referred to as MGP-CP^t) and the adaptive MGP (referred to as MGP-CP^a). For both methods, we use the Gaussian likelihood model for continuous data and the logistic model (with Pólya-Gamma sampling during inference) for binary data.

The following baselines were used for comparisons. (*i*) Bayesian CP (BCP), a fully Bayesian version of the standard probabilistic CP decomposition (Xiong et al., 2010). It assumes that the rank is known. (*ii*) ARD based CP (ARD-CP), a method that uses automatic relevance determination (ARD) (Mørup and Hansen, 2009; Zhao et al., 2014) to determine the rank of a tensor by inferring the relevant columns in the factor matrix of each mode. (*iii*) An Infinite Tucker Decomposition based on t process (InfTucker^{tp}), which is a kernel-based nonparametric Bayesian generalization of the low-rank Tucker decomposition(Xu et al., 2013), and is based on an *implicit* mapping of the component tensors to a higher (potentially infinite) dimensional space and performing a low-rank Tucker decomposition in that space. This method requires that the rank is given.

We evaluate our model and the various baselines on the following experiments: (*i*) tensor completion for continuous data, (*ii*) tensor completion for binary data, (*iii*) SVM based classification for EEG data using factors learned by different tensor decomposition methods, and (*iv*) image inpainting for color images by posing it as tensor completion problem for continuous data.

We initialize the MGP-CP^a using an initial rank = 1 and allow the rank to grow/shrink using our adaptation strategy discussed in Section 5.1.3. For MGP-CP^t and the ARD-CP baseline, we set the truncation level to a sufficiently large value. We run the sampling based methods for 1500 iterations with 1000 burn-in iterations, collect samples every 5 iterations after the burn-in phase, and report all results using the posterior sample based averages. For Bayesian CP and InfTucker^{tp}, which require

	Synthetic Data (R=10)	**Amino Acid**	**Flow-injection**	**EEG Data**
Bayesian CP	0.1231 ($\pm$0.0278)	**0.0004 ($\pm$0.0001)**	0.0012 ($\pm$0.0002)	0.1760 ($\pm$0.0032)
ARD-CP	0.0921 ($\pm$0.0006)	0.0350 ($\pm$0.0535)	0.0196 ($\pm$0.0133)	0.1860 ($\pm$0.0050)
InfTucker^{tp}	0.6644 ($\pm$0.0136)	0.8478 ($\pm$0.0103)	0.8382 ($\pm$0.0080)	0.5394 ($\pm$0.0823)
MGP-CP^{t}	**0.0922 ($\pm$0.0011)**	0.0006 ($\pm$0.0001)	0.0007 ($\pm$0.0003)	0.1622 ($\pm$0.0016)
MGP-CP^{a}	0.0935 ($\pm$0.0007)	0.0005 ($\pm$0.0001)	**0.0005 ($\pm$0.0003)**	**0.1608 ($\pm$0.0033)**

FIGURE 5.2: Continuous Data: MSE

the rank to be specified, we vary the ranks over a range and report the results using the rank that gave the best held-out data predictions.

5.4.1 Low-rank Tensor Completion: Continuous Data

We first experiment on the tensor completion task for continuous data. For this experiment, we use four datasets: (*i*) Synthetic data of size $20 \times 20 \times 20 \times 20$, generated as an equally-weighted sum of 10 rank-1 tensors of the same size (so the ground-truth rank is 10). (*ii*) Amino Acid data (Xu et al., 2013; Chu and Ghahramani, 2009) of size $5 \times 61 \times 201$, consisting of five laboratory-made amino acid samples. (*iii*) Flow Injection data (Xu et al., 2013; Chu and Ghahramani, 2009) of size $12 \times 100 \times 89$ obtained from a flow injection analysis (FIA) system. (*iv*) EEG data of size $15 \times 16 \times 560$ consisting of EEG measurements of 560 subjects. For this task, we treat 50% of the data as missing and reconstruct it using the model learned on the remaining 50% data. We report the results in terms of the mean-squared-error (MSE) on the reconstruction task. Each experiment is repeated 10 times with different splits of observed and missing data.

The results are shown in Figure 5.2. Both our models achieve reconstruction accuracies comparable to or better than the gold-standard Bayesian CP (which was given the ground-truth rank for synthetic data, and best rank chosen via held-out error on real-world datasets - 5 for Amino Acid, 6 for Flow-injection, 30 for EEG data). Moreover, on all datasets, both our models perform better than ARD-CP and InfTucker^{tp}.

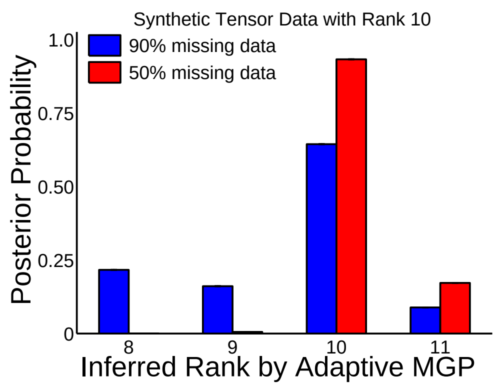


FIGURE 5.3: Empirical distribution of the inferred rank by MGP-CP^a run with 90% and 50% missing data (starting with $R = 1$)

To see whether our method can recover the true underlying rank, we run MGP-CP^a on the $20 \times 20 \times 20 \times 20$ synthetic data having a ground-truth rank 10, first with 90% missing data and then with 50% missing data. Figure 5.3 shows the posterior distribution of the inferred rank (based on the estimated empirical distribution of the ranks using posterior samples after the burn-in phase). As shown in the figure, in both cases, the posterior is concentrated at rank 10 and as the amount of training data increases from 90% missing to 50% missing, the posterior peaks further at rank 10. On real-world datasets, our method discovers ranks that are consistent with what is known from domain knowledge in the chemometrics literature on analyzing these datasets (our method infers the rank to be 3-4 on average on Amino Acid data and 6-7 on average on Flow Injection data).

5.4.2 Low-rank Tensor Completion: Binary Data

We next experiment with tensor completion for binary tensor data. We use four binary datasets for this experiment: (i) Synthetic data of size $20 \times 20 \times 20 \times 20$ having a ground-truth rank 10 (about 1.6% non-zero entries). (ii) Lazega-Lawyers multirelational social network data (Lazega, 2001) given in the form of a tensor of size $71 \times 71 \times 3$ (about 15% non-zero entries) containing three types of social networks (friendship, coworker, and advisory relationships) between 71 partners and associates in several New England law firms. (*iii*) Kinship multirelational data (Nickel et al., 2011) of size $104 \times 104 \times 26$ (about 3.84% non-zero entries) containing 26 types of kinship relations within the Alwayarra tribe. (*iv*) Nation multirelational data (Nickel et al., 2011) given in the form of a tensor of size $14 \times 14 \times 56$ (about 19% non-zero entries) containing 57 types of relationships (*e.g.*, export, protests, economic aid, etc.) among 14 countries. For each dataset, except Kinship, we treat 90% of the entries as missing and predict them using the rest 10% data. For Kinship data we use the experimental setting of 90% training and 10% test data as done in other recent works (Nickel et al., 2011; Jenatton et al., 2012). We use the area under the receiver-operating characteristic curve (AUC) score to compare the different methods in terms of their predictive ability. Each experiment is repeated 10 times with different splits of observed and missing data.

As the results in Figure 5.4 show, both our methods outperform the other baselines in terms of the AUC scores. It is noteworthy to see that on binary data the improvements of our methods over Bayesian CP are much more significant than the continuous-data case (even though the Bayesian CP baseline is provided with the ground-truth rank for the synthetic data and the best chosen rank based on held-out error for the real-world data). This can be attributed to the fact that Bayesian CP uses least-square minimization whereas our methods use the logistic loss. Because of this, for datasets having a significant number of zero entries (like the ones used in the experiments here), the Bayesian CP will tend to be biased towards predicting zeros. The ARD-CP baseline, although in principle able to infer the rank, suffers due to squared-loss minimization like Bayesian CP. InfTucker^{tp}, the next-best performing method, uses the logistic loss like our method; however, it relies on variational EM for inference and is prone to local-optimal issues (besides having to select the rank via cross-validation).

	Synthetic Data (R=10)	Lazega Lawyers	Kinship	Nation
Bayesian CP	0.6997 ($\pm$0.0434)	0.5671 ($\pm$0.0243)	0.9754 ($\pm$0.0022)	0.7230 ($\pm$0.0344)
ARD-CP	0.6045 ($\pm$0.0461)	0.5542 ($\pm$0.0378)	0.9842 ($\pm$0.0019)	0.6698 ($\pm$0.0527)
InfTucker^{tp}	0.8759 ($\pm$0.0143)	0.5982 ($\pm$0.0179)	0.9825 ($\pm$0.0022)	0.7981 ($\pm$0.0133)
MGP-CP^{t}	**0.9288 ($\pm$0.0140)**	0.6412 ($\pm$0.0101)	0.9896 ($\pm$0.0014)	**0.8105 ($\pm$0.0083)**
MGP-CP^{a}	0.9283 ($\pm$0.0109)	**0.6448 ($\pm$0.0139)**	**0.9909 ($\pm$0.0015)**	0.8096 ($\pm$0.0082)

FIGURE 5.4: Binary Data: AUC Scores

5.4.3 Binary Classification with Extracted Factors

We now experiment on an extrinsic evaluation task: binary classification using the factors learned via tensor decomposition. On the EEG data used in Section 5.4.1 for tensor completion experiments, we also have binary labels for each of the 560 subjects. We conduct an SVM based classification experiment where the factors extracted by various tensor decomposition methods are used to train an SVM. The tensor decomposition step uses only 50% of the total data and the 3^{rd} mode factors (the 3^{rd} mode represents the subjects) are used to train an SVM. After the tensor decomposition step, we use 10% of the subjects to train the SVM (using the extracted factors) and test on the remaining 90% subjects (to simulate a small sample size setting where a naïve approach of flattening the $15 \times 16 \times 560$ tensor a *matrix* could overfit). Because the tensor methods use only 50% data in the factor extraction stage, in the SVM experiment with the flattened tensor which resulted in a matrix of size 560×240, we hide 50% of the features and impute them using the respective feature means.

We repeat the classification experiment 20 times with different splits of training and test data. As Table 5.1 shows, the tensor decomposition methods perform better than SVM on flattened tensor which seems to overfit due to small sample size. Among the various tensor-decomposition-based methods, MGP-CP^{a} yields the best classification accuracy.

Table 5.1: Binary classification using factors learned from tensor decomposition

	Classification Accuracy
SVM (on flattened tensor)	65.02% (±3.10%)
Bayesian CP + SVM	67.95% (±4.39%)
ARD-CP + SVM	72.26% (±3.03%)
Infinite Tucker + SVM	66.53% (±3.32%)
MGP-CP^t + SVM	72.32% (±3.54%)
MGP-CP^a + SVM	**74.57% (±2.42%)**

5.4.4 Image Inpainting

Image inpainting is the task of completing an image with missing pixels. A two-dimensional RGB image can be treated as a three-dimensional tensor and the image inpainting task can be formulated as a tensor completion problem where the goal is to predict the values of the missing pixels using the observed pixel values. We apply our methods and the other baselines on this task using the benchmark Lena image of size $256 \times 256 \times 3$ for various fraction of missing pixels (90% missing, 80% missing, and 50% missing). Bayesian CP and InfTucker^{tp} were run with R ranging from 5 to 50 and we report the result with the best reconstruction error. For ARD-CP and MGP-CP^t, the truncation level was set to 50. The MGP-CP^a was initialized with $R = 1$. In Table 5.2, the reconstruction accuracies for each case are shown, and the reconstructed images for each case using our MGP-CP^t model are shown in Figure 5.5. As shown in Table 5.2, both MGP-CP^t and MGP-CP^a outperform the other baselines on this task in all the three cases. As Figure 5.5 shows, our method can recover the underlying ground-truth image up to a very reasonable quality even when the percentage of missingness is very high.

5.4.5 Scalability

To assess the scalability of our method, we run an experiment on a large but sparsely observed synthetic tensor dataset of size $1000 \times 1000 \times 1000$ having 1 billion cells but

FIGURE 5.5: Image Inpainting: **Top row:** Corrupted images with 90%, 80%, and 50% pixels missing. **Bottom row:** Reconstructed.

Table 5.2: Image Inpainting: Reconstruction errors (MSE) on different amounts (90%, 80%, and 50%) of missing pixels

	90%	**80%**	**50%**
Bayesian CP	0.0146	0.0099	0.0088
ARD-CP	0.0203	0.0197	0.0193
Infinite Tucker	0.2563	0.2106	0.1056
MGP-CP^t	0.0125	**0.0049**	**0.0023**
MGP-CP^a	**0.0102**	0.0057	0.0031

sparsely observed such that only 1 million entries are known. For this dataset, we vary the number of observations from 0.2 million to 1 million and run the MGP-CP^t with a fixed truncation level (so R and K stay fixed and only N varies) for 200 iterations in each case. Even when using an unoptimized MATLAB implementation, using our method we are able to deal with datasets of this scale in a reasonable amount of time. As shown in Figure 5.6, our method scales linearly with the number of observations.

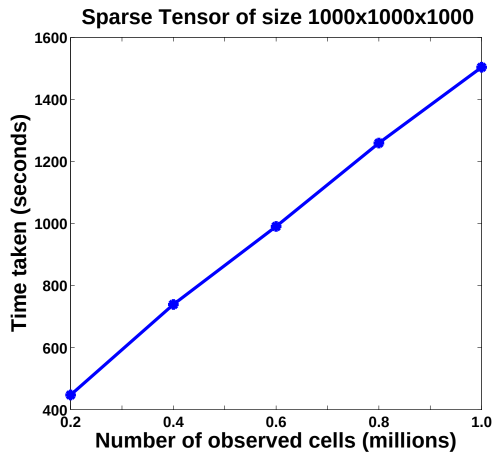


FIGURE 5.6: Linear scalability on a large-scale but sparse tensor

5.5 Summary

We have developed a flexible and scalable nonparametric Bayesian framework for analyzing multiway tensor data. Our framework is flexible as it does not require the tensor rank to be specified beforehand. The model can adapt its complexity (the rank of the decomposition which can grow or shrink as inference progresses) as apppropriate for the data under consideration. Our framework can naturally handle both continuous and binary datasets using suitable likelihood models. Bayesian inference can be efficiently done in both cases using closed-form Gibbs sampling which scales linearly in the number of observations in the tensor. The 2-way version of our model with binary observations can also be a scalable alternative to other state-of-the-art nonparametric Bayesian methods (Miller et al., 2009) for link-prediction in *single*-relational networks. Although in this work, we considered the CP decomposition with two specific likelihood models, integrating our tensor decomposition framework with other task-specific objectives (e.g., supervised classification or ranking for multiway data) could be another future avenue of work.

Appendix A

Lévy Measure Decomposition

A.1 Lévy measure decomposing of beta process

$$
\begin{aligned}
\nu(d\pi, d\omega) &= c(\omega)\pi^{-1}(1-\pi)^{c(\omega)-1} d\pi \mu(d\omega) \\
&= c(\omega) \sum_{k=0}^{\infty} (1-\pi)^k (1-\pi)^{c(\omega)-1} d\pi \mu(d\omega) \\
&= \sum_{k=0}^{\infty} \frac{\Gamma(c(\omega)+k+1)}{\Gamma(c(\omega)+k)\Gamma(1)} \pi^{1-1}(1-\pi)^{c(\omega)+k-1} d\pi \frac{c(\omega)}{c(\omega)+k} \mu(d\omega) \\
&= \sum_{k=0}^{\infty} \mathrm{Beta}(1, c(\omega)+k) d\pi \frac{c(\omega)}{c(\omega)+k} \mu(d\omega)
\end{aligned}
$$

A.2 Expectation of B_k

For $\forall \mathcal{A} \in \mathcal{F}$, denote $\mathbb{E}_{\mathrm{Poi}}$ and $\mathbb{E}_{\mathrm{Beta}}$ as the expectation computation incurred respectively by μ_k and $\mathrm{Beta}(1, c(\omega)+k)$; and $\{(\omega_{ki}, B_k(\omega_{ki}))\}_i$ a realization of Π_k. Apply restriction theorem and Campbell's theorem (Kingman, 1993):

$$\begin{aligned}\mathbb{E}(B_k(\mathcal{A})) &= \mathbb{E}_{\text{Poi}}(\mathbb{E}_{\text{Beta}}(B_k(\mathcal{A})))\\ &= \mathbb{E}_{\text{Poi}}\Big(\sum_{(\omega_{ki},\, B_k(\omega_{ki}))\,\in\,\Pi_k} \frac{1}{c(\omega_{ki})+k+1}\Big)\\ &= \int_{\mathcal{A}} \frac{1}{c(\omega)+k+1}\mu_k(d\omega)\end{aligned}$$

$$\begin{aligned}\sum_{k=0}^{\infty}\mathbb{E}(B_k(\mathcal{A})) &= \int_{\mathcal{A}}\sum_{k=0}^{\infty}\frac{1}{c(\omega)+k+1}\mu_k(d\omega)\\ &= \int_{\mathcal{A}}\sum_{k=0}^{\infty}\frac{c(\omega)}{[c(\omega)+k][c(\omega)+k+1]}\mu(d\omega)\\ &= \int_{\mathcal{A}}\sum_{k=0}^{\infty}c(\omega)\Big[\frac{1}{c(\omega)+k}-\frac{1}{c(\omega)+k+1}\Big]\mu(d\omega)\\ &= \int_{\mathcal{A}}\mu(d\omega) = \mu(\mathcal{A}) = \mathbb{E}(B(\mathcal{A}))\end{aligned}$$

A.3 Variance of B_k

For the variance, first calculate $\mathbb{E}(B_k^2(\mathcal{A}))$:

$$\begin{aligned}\mathbb{E}(B_k^2(\mathcal{A})) =& \mathbb{E}_{\text{Poi}}\Big(\mathbb{E}_{\text{Beta}}\Big(\sum_{(\omega_{ki},\, B_k(\omega_{ki}))\,\in\,\Pi_k} B_k^2(\omega_{ki}) + \sum_{\substack{(\omega_{ki},\, B_k(\omega_{ki}))\,\in\,\Pi_k\\ (\omega_{ki'},\, B_k(\omega_{ki'}))\,\in\,\Pi_k\\ \omega_{ki}\,\neq\,\omega_{ki'}}} B_k(\omega_{ki})B_k(\omega_{ki'})\Big)\Big)\\ =& \mathbb{E}_{\text{Poi}}\Big[\sum_{\omega_{ki},\, B_k(\omega_{ki}))\,\in\,\Pi_k} \frac{2}{(c(\omega_{ki})+k+1)(c(\omega_{ki})+k+2)}\\ &+ \sum_{\substack{(\omega_{ki},\, B_k(\omega_{ki}))\,\in\,\Pi_k\\ (\omega_{ki'},\, B_k(\omega_{ki'}))\,\in\,\Pi_k\\ \omega_{ki}\,\neq\,\omega_{ki'}}} \frac{1}{(c(\omega_{ki})+k+1)(c(\omega_{ki'})+k+1)}\Big]\end{aligned}$$

The first term $I_1 = \int_{\mathcal{A}} \frac{2}{(c(\omega)+k+1)(c(\omega)+k+2)} \mu_k(d\omega)$ by applying Campbell's theorem; For the second term I_2:

$$I_2 = \sum_{n=0}^{\infty} \frac{\exp(-\int_{\Omega} \mu_k(d\omega))(\int_{\Omega} \mu_k(d\omega))^n}{n!} \underbrace{\int_{\mathcal{A}} \cdots \int_{\mathcal{A}}}_{n} \Pi_{i=1}^{n} [\frac{\mu_k(d\omega_{ki})}{\int_{\Omega} \mu_k(d\omega)}] \cdot$$

$$\sum_{\substack{(\omega_{ki}, B_k(\omega_{ki})) \in \Pi_k \\ (\omega_{ki'}, B_k(\omega_{ki'})) \in \Pi_k \\ \omega_{ki} \neq \omega_{ki'}}} \frac{1}{(c(\omega_{ki}) + k + 1)(c(\omega_{ki'}) + k + 1)}$$

$$= \sum_{n=0}^{\infty} \frac{\exp(-\int_{\Omega} \mu_k(d\omega))(\int_{\Omega} \mu_k(d\omega))^n}{n!} (n^2 - n) \cdot$$

$$\int_{\mathcal{A}} \int_{\mathcal{A}} \frac{\mu_k(d\omega_{ki}) \mu_k(d\omega_{ki'})}{(\int_{\Omega} \mu_k(d\omega))^2} \frac{1}{(c(\omega_{ki}) + k + 1)(c(\omega_{ki'}) + k + 1)}$$

$$= (\int_{\mathcal{A}} \frac{1}{c(\omega) + k + 1} \mu_k(d\omega))^2$$

Thus: $\mathrm{Var}(B_k(\mathcal{A})) = \mathbb{E}(B_k^2(\mathcal{A})) - (\mathbb{E}(B_k(\mathcal{A})))^2 = \int_{\mathcal{A}} \frac{2}{(c(\omega)+k+1)(c(\omega)+k+2)} \mu_k(d\omega)$

$$\sum_{k=0}^{\infty} \mathrm{Var}(B_k(\mathcal{A})) = \int_{\mathcal{A}} \sum_{k=0}^{\infty} \frac{2c(\omega)}{(c(\omega) + k)(c(\omega) + k + 1)(c(\omega) + k + 2)} \mu(d\omega)$$

$$= \int_{\mathcal{A}} \sum_{k=0}^{\infty} c(\omega) [\frac{1}{c(\omega) + k} - \frac{2}{c(\omega) + k + 1} + \frac{1}{c(\omega) + k + 2}] \mu(d\omega)$$

$$= \int_{\mathcal{A}} \frac{1}{c(\omega) + 1} \mu(d\omega) = \mathrm{Var}(B(\mathcal{A}))$$

A.4 Truncation analysis of beta process

$$\begin{aligned}
\text{RHS of (15)} &= 1 - \mathbb{E}[\Pi_{k=K+1}^{\infty}\Pi_{i=1}^{n_k}(1-\pi_{ki})^M] \\
&\overset{(a)}{\leqslant} 1 - \{\Pi_{k=K+1}^{\infty}\mathbb{E}[\Pi_{i=1}^{n_k}(1-\pi_{ki})]\}^M \\
&= 1 - \{\Pi_{k=K+1}^{\infty}\mathbb{E}[\mathrm{e}^{\sum_{i=1}^{n_k}\log(1-\pi_{ki})}]\}^M \\
&\overset{(b)}{=} 1 - \{\Pi_{k=K+1}^{\infty}[\mathrm{e}^{\int_{\Omega\times(0,1)}\mathrm{e}^{\log(1-\pi)}\nu_k(d\pi,d\omega)}]\}^M \\
&= 1 - \mathrm{e}^{-M\int_\Omega\sum_{k=K+1}^{\infty}\frac{1}{c+K+1}\mu_k(d\omega)} \\
&= 1 - \mathrm{e}^{-M\int_\Omega \mu_{K+1}(d\omega)}
\end{aligned}$$

where (a) is justified by Jensen's inequality and (b) the Campbell's theorem.

By the Euler-Maclaurin formula,

$$\frac{\mathbb{E}(I_K)}{\gamma} = \sum_{k=0}^{K}\frac{1}{1+\frac{k}{c}} \approx c\cdot\log(1+\frac{K}{c}) + \frac{1}{2}(1+\frac{1}{1+\frac{K}{c}}) \overset{K\to\infty}{\approx} c\cdot\log(1+\frac{K}{c})$$

so $K \approx c(\mathrm{e}^{\frac{\mathbb{E}(I_K)}{c\gamma}}-1)$. Thus the $\mathcal{L}_1$ distance: $\frac{c}{c+K+1} \approx \mathrm{e}^{-\frac{\mathbb{E}(I_K)}{c\gamma}}$. For the stick-breaking construction of (Paisley et al., 2010), $(\frac{c}{c+1})^{K+1} \approx (1+\frac{1}{c})^{-\frac{\mathbb{E}(I_K)}{\gamma}}$. With $c\to\infty$, $(1+\frac{1}{c})^c \uparrow \mathrm{e}$, thus $\mathrm{e}^{-\frac{\mathbb{E}(I)}{c\gamma}} < (1+\frac{1}{c})^{-\frac{\mathbb{E}(I)}{\gamma}}$.

A.5 Limit of the Lévy measure of IBP

Since $\nu_{\text{IBP}} = \frac{N}{\gamma}\text{Beta}(c\frac{\gamma}{N},c)d\pi\mu(d\omega) = \frac{N}{\gamma\Gamma(c\frac{\gamma}{N})}\pi^{c\frac{\gamma}{N}-1}(1-\pi)^{c-1}d\pi\mu(d\omega)$, to prove (22) is equal to prove $\frac{\gamma}{N}\Gamma(c\frac{\gamma}{N}) \overset{N\to\infty}{=} \frac{1}{c}$:

$$
\begin{aligned}
\frac{\gamma}{N}\Gamma(c\frac{\gamma}{N}) &= \Delta \int_0^\infty \mathrm{e}^{-t} t^{c\Delta-1} dt, \quad \text{with } \Delta = \frac{\gamma}{N} \\
&= \frac{1}{c}\int_0^\infty \mathrm{e}^{-t}(c\Delta t^{c\Delta-1})dt \\
&= \frac{1}{c}(\mathrm{e}^{-t}t^{c\Delta}|_0^\infty - \int_0^\infty -\mathrm{e}^{-t}t^{c\Delta}dt) \\
&\overset{N\to\infty}{=} \frac{1}{c}
\end{aligned}
$$

A.6 Lévy measure decomposing of gamma process

For the Lévy measure of the gamma process $\nu(dp, d\omega) = p^{-1}\mathrm{e}^{-\frac{p}{\theta(\omega)}}dp\alpha(d\omega)$, decompose the exponential part into: $\mathrm{e}^{-\frac{p}{\theta(\omega)}} = \mathrm{e}^{\frac{p}{\theta(\omega)}}\mathrm{e}^{-\frac{p}{\theta(\omega)/2}}$, and apply Taylor series expansion on $\mathrm{e}^{\frac{p}{\theta(\omega)}}$:

$$
\begin{aligned}
\nu(dp, d\omega) &= p^{-1}[\sum_{h=0}^{\infty}\frac{(\frac{p}{\theta})^h}{h!}]\mathrm{e}^{-\frac{p}{\theta/2}}dp\alpha(d\omega) \\
&= [p^{-1} + \sum_{h=1}^{\infty}\frac{p^{h-1}}{\theta^h h!}]\mathrm{e}^{-\frac{p}{\theta/2}}dp\alpha(d\omega) \\
&= \sum_{h=1}^{\infty}\mathrm{Gamma}(h, \theta/2)dp\frac{\alpha(d\omega)}{2^h h} + p^{-1}\mathrm{e}^{-\frac{p}{\theta/2}}dp\alpha(d\omega)
\end{aligned}
$$

thus $G = \Gamma_1 + \Gamma\mathrm{P}(\alpha, \theta(\omega)/2)$. Further decompose the exponential part of the gamma process $\Gamma\mathrm{P}(\alpha, \theta(\omega)/2)$ yields $G = \Gamma_1 + \Gamma_2 + \Gamma\mathrm{P}(\alpha, \theta(\omega)/3)$. Keep on this manipulation:

$$G = \sum_{k=1}^{\infty} \Gamma_k, \quad \Gamma_k = \sum_{h=1}^{\infty} \Gamma_{kh}$$

$$\nu_{kh} = \text{Gamma}(h, \frac{\theta}{k+1}) dp \frac{\alpha(d\omega)}{(k+1)^h h}$$

with Γ_{kh} a Lévy process with ν_{kh} its Lévy measure. Here $\text{Gamma}(h, \frac{\theta}{k+1})$ is the PDF of Gamma distribution with shape h and scale $\frac{\theta}{k+1}$.

A.7 The expectation of Γ_k and Γ_{kh}

For $\forall \mathcal{A} \in \mathcal{F}$, with Campbell's theorem applied,

$$\mathbb{E}(\Gamma_{kh}(\mathcal{A})) = \int_{\mathcal{A}} \frac{h\theta(\omega)}{k+1} \frac{\alpha(d\omega)}{(k+1)^h h} = \frac{\int_{\mathcal{A}} \theta(\omega)\alpha(d\omega)}{(k+1)^{h+1}}$$

$$\mathbb{E}(\Gamma_k(\mathcal{A})) = \sum_{h=1}^{\infty} \mathbb{E}(\Gamma_{kh}(\mathcal{A})) = \frac{\int_{\mathcal{A}} \theta(\omega)\alpha(d\omega)}{k(k+1)}$$

$$\mathbb{E}(G(\mathcal{A})) = \sum_{k=1}^{\infty} \mathbb{E}(\Gamma_k(\mathcal{A})) = \int_{\mathcal{A}} \theta(\omega)\alpha(d\omega)$$

A.8 The variance of Γ_k and Γ_{kh}

The variance of Γ_{kh} can be calculated with the method described in Section A.3:

$$\text{Var}(\Gamma_{kh}(\mathcal{A})) = \int_{\mathcal{A}} \frac{(h+1)\theta^2(\omega)}{(k+1)^{h+2}} \alpha(d\omega)$$

$$
\begin{aligned}
\mathrm{Var}(\Gamma_k(\mathcal{A})) &= \sum_{h=1}^{\infty} \mathrm{Var}(\Gamma_{kh}(\mathcal{A})) \\
&= \sum_{h=1}^{\infty} \int_{\mathcal{A}} \frac{(h+1)\theta^2(\omega)}{(k+1)^{h+2}} \alpha(d\omega) \\
&= \int_{\mathcal{A}} [\sum_{h=1}^{\infty} -\frac{\theta^2(\omega)}{(k+1)^{(h+1)}}]'_k \alpha(d\omega) \\
&= [\frac{1}{k^2} - \frac{1}{(k+1)^2}] \int_{\mathcal{A}} \theta^2(\omega)\alpha(d\omega)
\end{aligned}
$$

$$
\mathrm{Var}(G(\mathcal{A})) = \sum_{k=1}^{\infty} \mathrm{Var}(\Gamma_k(\mathcal{A})) = \int_{\mathcal{A}} \theta^2(\omega)\alpha(d\omega)
$$

Appendix B

Kernel beta process

B.1 Proof of Theorem 1

We prove (6) is the Lévy measure of the KBP, as defined in Theorem 1, following the same method with which the Lévy measure of the beta process is derived in Theorem 3.1 in (Hjort, 1990). We have the following notation: $A_{n,m}$ is the m^{th} part of the n-equipartition of Ω, $\omega_{n,m}$ is the central point of $A_{n,m}$, $j = \sqrt{-1}$, and $\boldsymbol{u} \in \mathbb{R}^{|\mathcal{S}|}$. The proof proceeds with the following sequence of steps.

$$
\begin{aligned}
&\mathbb{E}\{e^{j<\boldsymbol{u},\, \boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{A}})>}\} \\
&= \mathbb{E}\{e^{\int_{\mathcal{A}} j<\boldsymbol{u},\, \boldsymbol{\mathcal{B}}(\boldsymbol{d\omega})>}\} \\
&\overset{n\to\infty}{=} \mathbb{E}\{e^{\sum_{A_{n,m}\in\mathcal{A}} j<\boldsymbol{u},\boldsymbol{K_{n,m}}\pi_{n,m}>}\} \\
&= \mathbb{E}\{\Pi_{A_{n,m}\in\mathcal{A}} e^{j<\boldsymbol{u},\boldsymbol{K_{n,m}}\pi_{n,m}>}\} \\
&\overset{(\mathscr{A})}{=} \Pi_{A_{n,m}\in\mathcal{A}}\, \mathbb{E}\{e^{j<\boldsymbol{u},\boldsymbol{K_{n,m}}>\pi_{n,m}}\} \\
&= e^{\sum_{A_{n,m}\in\mathcal{A}} \log\mathbb{E}\{e^{j<\boldsymbol{u},\boldsymbol{K_{n,m}}>\pi_{n,m}}\}} \\
&\overset{(\mathscr{B})}{=} e^{\sum_{A_{n,m}\in\mathcal{A}}\, \mathbb{E}\{\sum_{k=1}^{\infty} \frac{[j<\boldsymbol{u},\boldsymbol{K_{n,m}}>\pi_{n,m}]^k}{k!}\}} \\
&\overset{(\mathscr{C})}{=} e^{\sum_{A_{n,m}\in\mathcal{A}} \sum_{k=1}^{\infty} \mathbb{E}\{\frac{[j<\boldsymbol{u},\boldsymbol{K_{n,m}}>]^k}{k!}\}\mathbb{E}\{[\pi_{n,m}]^k\}} \\
&\overset{(\mathscr{D})}{=} e^{\sum_{A_{n,m}\in\mathcal{A}} \sum_{k=1}^{\infty} \mathbb{E}\{\frac{[j<\boldsymbol{u},\boldsymbol{K_{n,m}}>]^k}{k!}\}\int_0^1 [\pi_{n,m}]^k\cdot \mathrm{Beta}(c(\omega_{n,m})B_0(A_{n,m}),c(\omega_{n,m})(1-B_0(A_{n,m})))d\pi_{n,m}} \\
&= e^{\sum_{A_{n,m}\in\mathcal{A}} \sum_{k=1}^{\infty} \mathbb{E}\{\frac{[j<\boldsymbol{u},\boldsymbol{K_{n,m}}>]^k}{k!}\}\frac{[c(\omega_{n,m})B_0(A_{n,m})+k-1]\cdots[c(\omega_{n,m})B_0(A_{n,m})+1]}{[c(\omega_{n,m})+k-1]\cdots[c(\omega_{n,m})+1]}B_0(A_{n,m})} \\
&\overset{n\to\infty}{=} e^{\sum_{A_{n,m}\in\mathcal{A}} \sum_{k=1}^{\infty} \mathbb{E}\{\frac{[j<\boldsymbol{u},\boldsymbol{K_{n,m}}>]^k}{k!}\}\frac{(k-1)\cdots 2\cdot 1}{[c(\omega)+k-1]\cdots[c(\omega)+1]}B_0(d\omega)} \\
&= e^{\int_{\mathcal{X}\times\Psi\times[0,1]\times\mathcal{A}} \sum_{k=1}^{\infty} \frac{[j<\boldsymbol{u},\boldsymbol{K}>]^k}{k!}\pi^k H(dx^*)Q(d\psi^*)c(\omega)\pi^{-1}(1-\pi)^{c(\omega)-1}d\pi B_0(d\omega)} \\
&= e^{\int_{\mathcal{X}\times\Psi\times[0,1]\times\mathcal{A}} \sum_{k=1}^{\infty} \frac{[j<\boldsymbol{u},\boldsymbol{K}\pi>]^k}{k!}\nu_{\mathcal{X}}(dx^*,d\psi^*,d\pi,d\omega)} \\
&\overset{(\mathscr{E})}{=} e^{\int_{\mathcal{X}\times\Psi\times[0,1]\times\mathcal{A}}(e^{j<\boldsymbol{u},\boldsymbol{K}\pi>}-1)\nu_{\mathcal{X}}(dx^*,d\psi^*,d\pi,d\omega)}
\end{aligned}
$$

where the steps of the proof are justified as follows. $(\mathscr{A})$: $\boldsymbol{K_{n,m}}$ and $\pi_{n,m}$ are independent on disjoint sets $\{A_{n,m}\}_{m=1}^{n}$; $(\mathscr{B})$: $\log(\mathbb{E}\{e^{j<\boldsymbol{u},\boldsymbol{K_{n,m}}>\pi_{n,m}}-1\}+1) = \mathbb{E}\{e^{j<\boldsymbol{u},\boldsymbol{K_{n,m}}>\pi_{n,m}}-1\}$ since $\mathbb{E}\{e^{j<\boldsymbol{u},\boldsymbol{K_{n,m}}>\pi_{n,m}}-1\}$ is infinitesimal. By Taylor series expansion: $\mathbb{E}\{e^{j<\boldsymbol{u},\boldsymbol{K_{n,m}}>\pi_{n,m}}-1\} = \mathbb{E}\{\sum_{k=1}^{\infty}\frac{[j<\boldsymbol{u},\boldsymbol{K_{n,m}}>\pi_{n,m}]^k}{k!}\}$. $(\mathscr{C})$: $\boldsymbol{K_{n,m}}$ and $\pi_{n,m}$ are independent to each other. $(\mathscr{D})$: By the definition of the beta process:

$\pi_{n,m} \sim \text{Beta}(c(\omega_{n,m})B_0(A_{n,m}), c(\omega_{n,m})(1 - B_0(A_{n,m})))$. $(\mathscr{E})$: By Taylor series expansion backward.

B.2 Properties of the KBP

Let $\mathcal{B}_x$, $\mathcal{B}_{x'}$ $(x \neq x')$ be a draw of KBP at covariate x and x', for $\forall \mathcal{A} \subset \mathcal{F}$:

$$\begin{aligned}
\mathbb{E}\mathcal{B}_x(\mathcal{A}) &= \mathbb{E}\int_{\mathcal{A}} \mathcal{B}_x(d\omega) = \int_{\mathcal{A}} \mathbb{E}\mathcal{B}_x(d\omega) = \int_{\mathcal{A}} \mathbb{E}(K_x B(d\omega)) \\
&= \int_{\mathcal{A}} \mathbb{E}K_x \mathbb{E}B(d\omega) = \int_{\mathcal{A}} B_0(d\omega)\mathbb{E}K_x = B_0(\mathcal{A})\mathbb{E}K_x \\
\text{Cov}(\mathcal{B}_x(\mathcal{A}), \mathcal{B}_{x'}(\mathcal{A})) &= \int_{\mathcal{A}} \text{Cov}(\mathcal{B}_x(d\omega), \mathcal{B}_{x'}(d\omega)) = \int_{\mathcal{A}} \text{Cov}(K_x B(d\omega), K_{x'} B(d\omega)) \\
&= \int_{\mathcal{A}} \mathbb{E}(K_x K_{x'})\mathbb{E}(B^2(d\omega)) - \mathbb{E}K_x \mathbb{E}K_{x'} B_0^2(d\omega) \\
&= \mathbb{E}(K_x K_{x'}) \int_{\mathcal{A}} \frac{B_0(d\omega)(1 - B_0(d\omega))}{c(\omega) + 1} - \text{Cov}(K_x, K_{x'}) \int_{\mathcal{A}} B_0^2(d\omega)
\end{aligned}$$

We are especially interested in the conditional correlation between $\mathcal{B}_x(\mathcal{A})$ and $\mathcal{B}_{x'}(\mathcal{A})$ when kernel parameters $\{x_i^*, \psi_i^*\}_{i=1}^{\infty}$ are given. Denote the mass parameter $\gamma = B_0(\Omega)$. In practice we truncate the number of terms used in (7) to I, then we have the expectation of π_i: $\alpha = \gamma/I$ with $\pi_i \sim \text{Beta}(c\alpha, c(1-\alpha))$ for $i = 1, 2, \cdots, I$. Denote $\boldsymbol{K_x} = (K(x, x_1^*, \psi_1^*), \cdots, K(x, x_i^*, \psi_i^*), \cdots)^T$ with $i : \omega_i \in \mathcal{A}$. Hence:

$$
\begin{aligned}
\operatorname{Corr}(\mathcal{B}_x(\mathcal{A}), \mathcal{B}_{x'}(\mathcal{A})) &= \frac{\int_{\mathcal{A}} \operatorname{Cov}(K_x B(d\omega), K_{x'} B(d\omega))}{\{\int_{\mathcal{A}} \operatorname{Var}(K_x B(d\omega)) \cdot \int_{\mathcal{A}} \operatorname{Var}(K_{x'} B(d\omega))\}^{\frac{1}{2}}} \\
&= \frac{\sum_{i:\omega_i \in \mathcal{A}} K(x, x_i^*, \psi_i^*) K(x', x_i^*, \psi_i^*) \operatorname{Var}(\pi_i)}{\{\sum_{i:\omega_i \in \mathcal{A}} K^2(x, x_i^*, \psi_i^*) \operatorname{Var}(\pi_i) \cdot \sum_{i:\omega_i \in \mathcal{A}} K^2(x', x_i^*, \psi_i^*) \operatorname{Var}(\pi_i)\}^{\frac{1}{2}}} \\
&= \frac{\frac{\alpha(1-\alpha)}{c+1} \sum_{i:\omega_i \in \mathcal{A}} K(x, x_i^*, \psi_i^*) K(x', x_i^*, \psi_i^*)}{\frac{\alpha(1-\alpha)}{c+1} \{\sum_{i:\omega_i \in \mathcal{A}} K^2(x, x_i^*, \psi_i^*) \cdot \sum_{i:\omega_i \in \mathcal{A}} K^2(x', x_i^*, \psi_i^*)\}^{\frac{1}{2}}} \\
&= \frac{\sum_{i:\omega_i \in \mathcal{A}} K(x, x_i^*, \psi_i^*) K(x', x_i^*, \psi_i^*)}{\{\sum_{i:\omega_i \in \mathcal{A}} K^2(x, x_i^*, \psi_i^*) \cdot \sum_{i:\omega_i \in \mathcal{A}} K^2(x', x_i^*, \psi_i^*)\}^{\frac{1}{2}}} \\
&= \frac{< \boldsymbol{K_x}, \boldsymbol{K_{x'}} >}{\| \boldsymbol{K_x} \|_2 \cdot \| \boldsymbol{K_{x'}} \|_2}
\end{aligned}
$$

Bibliography

Applebaum, D. (2009), *Levy Processes and Stochastic Calculus*, Cambridge University Press.

Bazerque, J., Mateos, G., and Giannakis, G. (2013), "Rank regularization and Bayesian inference for tensor completion and extrapolation," *arXiv preprint arXiv:1301.7619*.

Bhattacharya, A. and Dunson, D. (2011), "Sparse Bayesian infinite factor models," *Biometrika*.

Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.

Billingsley, P. (1995), *Probability and Measure*, Wiley-Interscience, 3 edn.

Blei, D., Griffiths, T., and Jordan, M. (2010), "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *J. ACM*, 57.

Bordes, A., Glorot, X., Weston, J., and Bengio, Y. (2012), "A semantic matching energy function for learning with multi-relational data," *Machine Learning*.

Brix, A. (1999), "Generalized Gamma Measures and Shot-Noise Cox Processes," *Advances in Applied Probability*, 31, 929–953.

Broderick, T., Jordan, M., and Pitman, J. (2011), "Beta processes, stick-breaking, and power laws," *Bayesian analysis*.

Chu, W. and Ghahramani, Z. (2009), "Probabilistic Models for Incomplete Multidimensional Arrays," in *AISTATS*.

Çinlar, E. (2010), *Probability and Stochastics*, Graduate Texts in Mathematics, Springer.

de Silva, V. and Lim, L. (2008), "Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem," *SIAM J. Matrix Analysis Applications*.

Doshi, F., Miller, K., Van Gael, J., and Teh, Y. (2009), "Variational Inference for the Indian Buffet Process," in *AISTATS*, vol. 12.

Dunson, D. and Xing, C. (2012), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *JASA*.

Dunson, D. B. and Park, J.-H. (2008), "Kernel stick-breaking processes," *Biometrika*, 95, 307–323.

Eckart, C. and Young, G. (1936), "The approximation of one matrix by another of lower rank," *Psychometrika*.

Ferguson, T. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209–230.

Ferguson, T. and Klass, M. (1972), "A Representation of Independent Increment Processes without Gaussian Components," *The Annals of Mathematical Statistics*.

Foti, N., Futoma, J., Rockmore, D., and Williamson, S. (2013), "A unifying representation for a class of dependent random measures," in *AISTATS*.

Griffiths, T. and Ghahramani, Z. (2005), "Infinite latent feature models and the Indian buffet process," in *NIPS*.

Griffiths, T. L. and Ghahramani, Z. (2011), "The indian buffet process: An introduction and review," *JMLR*.

Hastad, J. (1990), "Tensor rank is NP-complete," *Journal of Algorithms*.

Hjort, N. (1990), "Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data," *Annals of Statistics*.

Hjort, N., Holmes, C., Mueller, P., and Walker, S. (2010), *Bayesian Nonparametrics: Principles and Practice*, Cambridge University Press, Cambridge, UK.

Jenatton, R., Le Roux, N., Bordes, A., and Obozinski, G. (2012), "A latent factor model for highly multi-relational data," in *NIPS*.

Jordan, M. (2009), "Hierarchical models, nested models and completely random measures," in *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*, New York: Springer.

Kingman, J. (1967), "Completely Random Measure," in *Pacific Journal of Mathematics*, vol. 21(1):59-78.

Kingman, J. (1993), *Poisson Processes*, Oxford University Press, Oxford.

Kingman, J. (2002), *Poisson Processes*, Oxford Press.

Knowles, D. and Ghahramani, Z. (2007), "Infinite Sparse Factor Analysis and Infinite Independent Components Analysis," in *Independent Component Analysis and Signal Separation*.

Kolda, T. G. and Bader, B. W. (2009), "Tensor decompositions and applications," *SIAM review*.

Lazega, E. (2001), *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*, Oxford University Press on Demand.

Lijoi, A. and Prünster, I. (2014), "Bayesian inference with dependent normalized completely random measures," *Bernoulli*.

Lin, D., Grimson, E., and Fisher, J. (2010), "Construction of Dependent Dirichlet Processes based on Poisson Processes," in *NIPS*, pp. 1396–1404.

MacEachern, S. (1999), "Dependent Nonparametric Processes," in *In Proceedings of the Section on Bayesian Statistical Science*.

Miller, K., Griffiths, T., and Jordan., M. I. (2008), "The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features," in *UAI*.

Miller, K., Jordan, M. I., and Griffiths, T. L. (2009), "Nonparametric latent feature models for link prediction," in *NIPS*.

Mørup, M. and Hansen, L. K. (2009), "Automatic relevance determination for multi-way models," *Journal of Chemometrics*.

Nickel, M. and Tresp, V. (2013), "Logistic Tensor Factorization for Multi-Relational Data," *arXiv preprint arXiv:1306.2084*.

Nickel, M., Tresp, V., and Kriegel, H. (2011), "A three-way model for collective learning on multi-relational data," in *ICML*.

Paisley, J. and Carin, L. (2009), "Nonparametric Factor Analysis with Beta Process Priors," in *ICML*.

Paisley, J., Zaas, K., Woods, C., Ginsburg, G., and Carin, L. (2010), "A Stick-Breaking Construction of the Beta Process." in *ICML*, pp. 847–854.

Paisley, J., Blei, D., and Jordan, M. (2012), "Stick-breaking beta processes and the Poisson process," *AISTATS*.

Pitman, J. (1995), "Exchangeable and Partially Exchangeable Random Partitions," *Probability Theory and Related Fields*.

Polson, N., Scott, J., and Windle, J. (2012), "Bayesian inference for logistic models using Polya-Gamma latent variables, http://arxiv.org/abs/1205.0310," .

Rao, V. and Teh, Y. (2009), "Spatial Normalized Gamma Processes," in *NIPS*.

Rasmussen, C. and Williams, C. (2006), *Gaussian Processes for Machine Learning*, MIT Press.

Ren, L., Dunson, D., Lindroth, S., and Carin, L. (2010), "Dynamic Nonparametric Bayesian Models for Analysis of Music," *Journal of The American Statistical Association*, 105, 458–472.

Ren, L., Wang, Y., Dunson, D., and Carin, L. (2011a), "The Kernel Beta Process," in *NIPS*.

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011b), "Logistic Stick-Breaking Process," *J. Machine Learning Research*.

Rodriguez, A. and Dunson, D. B. (2009), "Nonparametric Bayesian models through probit stickbreaking processes," *Univ. California Santa Cruz Technical Report*.

Sato, K. (1999), *Lévy processes and infinitely divisible distributions*, Cambridge University Press.

Sethuraman, J. (1994a), "A constructive definition of Dirichlet priors," *Statistica Sinica*.

Sethuraman, J. (1994b), "A constructive definition of Dirichlet priors," *Statistica Sinica*.

Teh, Y. (2006), "A Hierarchical Bayesian Language Model based on Pitman-Yor Processes," in *Coling/ACL*, pp. 985–992.

Teh, Y. and Görür, D. (2009), "Indian Buffet Processes with Power-law Behavior," in *NIPS*.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006), "Hierarchical Dirichlet processes," *JASA*.

Teh, Y., Görür, D., and Ghahramani, Z. (2007), "Stick-breaking construction for the Indian buffet process," in *AISTATS*.

Thibaux, R. (2008), "Nonparametric Bayesian Models for Machine Learning," Ph.D. thesis, EECS Dept., University of California, Berkeley.

Thibaux, R. and Jordan, M. (2007a), "Hierarchical beta processes and the Indian buffet process," in *AISTATS*.

Thibaux, R. and Jordan, M. I. (2007b), "Hierarchical beta processes and the Indian buffet process," in *AISTATS*.

Tomioka, R., Hayashi, K., and Kashima, H. (2010), "Estimation of low-rank tensors via convex optimization," *arXiv preprint arXiv:1010.0789*.

Wang, Y. and Carin, L. (2012), “Levy Measure Decompositions for the Beta and Gamma Processes,” in *ICML*.

Williamson, S., Orbanz, P., and Ghahramani, Z. (2010), “Dependent Indian buffet processes,” in *AISTATS*.

Wolpert, R., Clyde, M., and Tu, C. (2011), “Stochastic Expansions using Continuous Dictionaries: Lévy Adaptive Regression Kernels,” *Annals of Statistics*.

Woods, C., McClain, M., Chen, M., Zaas, A., Nicholson, B., Varkey, J., Veldman, T., Kingsmore, S., Huang, Y., Lambkin-Williams, R., Gilbert, A., Ramsburg, A. H. E., Glickman, S., Lucas, J., Carin, L., and Ginsburg, G. (2013), “A Host Transcriptional Signature for Presymptomatic Detection of Infection in Humans Exposed to Influenza H1N1 or H3N2,” *PLoS ONE*.

Xiong, L., Chen, X., Huang, T., Schneider, J. G., and Carbonell, J. G. (2010), “Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization.” in *SDM*.

Xu, Z., Yan, F., and Qi, Y. (2013), “Bayesian Nonparametric Models for Multiway Data Analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yoshii, K., Tomioka, R., Mochihashi, D., and Goto, M. (2013), “Infinite Positive Semidefinite Tensor Factorization for Source Separation of Mixture Signals,” in *ICML*.

Yu, K., Chu, W., Yu, S., Tresp, V., and Xu, Z. (2007), “Stochastic relational models for discriminative link prediction,” in *NIPS*.

Zhao, Q., Zhang, L., and Cichocki, A. (2014), “Bayesian CP Factorization of Incomplete Tensors with Automatic Rank Determination,” *arXiv preprint arXiv:1401.6497*.

Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., and Carin, L. (2009), “Non-Parametric Bayesian Dictionary Learning for Sparse Image Representations,” in *NIPS*.

Zhou, M., Yang, H., Sapiro, G., Dunson, D., and Carin, L. (2011), “Dependent Hierarchical Beta Process for Image Interpolation and Denoising,” in *AISTATS*.

Biography

Yingjian Wang was born in Chaoyang, Liaoning Province, in the northeastern China. He came into the PhD program at the Electrical and Computer Engineering Department of Duke University in the summer 2009. Yingjian's research interest is around stochastic processes and their applications in Machine Learning. Before Yingjian came to Duke, he had graduated from Tsinghua University with a Master of Engineering degree and Beijing Institute of Technology with a Bachelor of Engineering degree, both in the Department of Electronic Engineering.

The people adored most by Yingjian are his parents - Zongzhen Wang and Guifen Li. There are three places in the world Yingjian cherishes most: the plaza around the Pagoda in his hometown Chaoyang, the Lotus Pond of the Tsinghua University, and the Sarah P. Duke Garden of the Duke University.